THE UNIVERSITY
of EDINBURGH

Genetic diversity in the processing and transcriptomic diversity in the targeting of microRNAs

Jonathan Moody

THE UNIVERSITY
of EDINBURGH

Doctor of Philosophy
The University of Edinburgh
2016

Declaration

I hereby declare that this thesis has been composed solely by myself and has not been accepted in any previous candidature for a higher degree. All work presented in this thesis was, unless acknowledged, initiated and executed by myself. All sources in the text have been acknowledged by reference.

Jonathan Moody
September 2016

Abstract

MicroRNAs are short RNA molecules that are central to the regulation of many cellular and developmental pathways. They are processed in several stages from structured precursors in the nucleus, into mature microRNAs in the cytoplasm where they direct protein complexes to regulate gene expression through, often imperfect base-pairing with target messenger RNAs. The broad aim of this project is to better understand how polymorphisms and new mutations can disrupt microRNA processing and targeting, and ultimately define their contributions to human disease.

I have taken two approaches towards this. The first approach is to comprehensively identify the microRNA targets by developing and applying a novel computational pipeline to identify microRNA binding events genome-wide in RNA-RNA interaction datasets. I use this to examine the transcriptomic diversity of microRNA binding, finding microRNA binding events along the full length of protein coding transcripts and with a variety of non-coding RNAs. This reveals enrichment for non-canonical microRNA binding at promoters and intronic regions around splice sites, and identifies highly spatially clustered binding sites within transcripts that may be acting as competitive endogenous RNAs to compete for microRNAs, effectively sequestering them. Using statistical models and new cell fractionated RNA-seq data, I rank the features of microRNAs and their binding sites which contribute to the strength and specificity of their interaction to provide a better understanding of the major determinants of microRNA targeting.

The second approach is to directly identify DNA sequence changes in microRNA precursors that alter processing efficiency affecting mature microRNA abundance which are routinely overlooked in the search for disease or trait associated causal variants. I have systematically screened public datasets for both rare and common polymorphisms that overlap microRNA precursors and are correlated with mature microRNA levels as measured in short RNA sequencing. I use these eQTL SNPs to examine the most important microRNA precursor regions and sequence motifs. Several of these SNPs have been observed as risk factors in cancer or other clinically relevant traits, and correlated with microRNA processing efficiency. I demonstrate that a specific DNA change which is known to be important in the development of some cancers, is located in a microRNA precursor and affects the balance of its two products, miR-146a-3p and miR-146a-5p, that can be produced from that single precursor providing

new insights into the mechanisms of microRNA production and the aspects of genetic mis-regulation that result in cancer. I find further examples of common human polymorphisms that appear to affect microRNA production from their precursors, several of these variants are independently implicated in human immune disease, cancer susceptibility and associated with other complex traits. As they exhibit a molecular phenotype and immediately lead to mechanistic hypotheses of trait causality that can be tested, these variants could provide a route into the frequently intractable problem of mechanistically linking non-coding genetic variation to human phenotypes. Applying similar studies to patient DNA has revealed rare and unique DNA changes that are now candidates for causing human disease that are being subject to follow-up experimental studies. Collectively this work has started to define which sequences changes in microRNAs are likely to disrupt their function and provides a paradigm for the analysis of microRNA sequence variants in human genetic disease.

Lay summary

Most cells of the human body have essentially the same genetic blueprint in their DNA. What makes a skin cell different from a neuron and both different from a muscle cell are the distinct ranges and amounts of products that are made from that blueprint. Careful regulation of the dosage of those products, often proteins, is central to health and development, and disrupted regulation can result in disease and is the molecular basis of cancer. Small RNA molecules termed microRNAs are known to be a key component of the dosage regulation system. microRNAs are produced from longer intricately folded precursor RNAs and work by targeting proteins to longer RNA molecules containing sequences related to the microRNA, often causing those targets to be broken down.

The broad aim of this project is to better understand how DNA sequence changes can disrupt microRNAs and ultimately contribute to human disease. I have taken two approaches to this, the first to comprehensively define what the microRNA targets are and how those targets are found by the microRNA, the second approach is to directly identify DNA sequence changes that alter the amount of microRNA produced. Together these approaches can implicate how changes in a microRNA or its production can have knock-on effects.

I have developed new computational tools to analyse high-throughput data showing microRNA interactions. For the first time this allows the whole genome scale search for targets without making assumptions about the type of target. This has revealed several new types of microRNA target but also confirmed the importance of the previously known type of target. This approach, coupled with some new experimental data also allowed me to measure the specificity of interactions and where within a cell they are taking place.

Looking at how changes in the DNA blueprint for microRNAs affects their production from precursors, I have found some DNA changes causing either more or less microRNA to be produced. I demonstrate that a specific DNA change which is known to be important in the development of some cancers, is located in a microRNA precursor and affects the balance of two products that can be produced from that precursor. This provides new insights into the mechanisms of microRNA production and the genetic mis-regulation that results in cancer. I also find further DNA changes that appear to affect microRNA production from their precursors, several of these changes are common in the human population and also seem to correspond to clinical traits such as blood lipid levels. Applying similar studies to patient DNA

has revealed rare and unique DNA changes that are now candidates for causing human disease. Collectively this work has started to define which sequence changes in microRNAs are likely to disrupt their function and provides a paradigm for the analysis of microRNA sequence variants in human genetic disease.

Acknowledgements

With thanks to my supervisors Martin Taylor and Greg Kudla for their supervision and support, members of my thesis committee Colin Semple and Javier Caceres for their useful opinions.

Thanks to members of the evogen group for their encouragement and helpful discussions.

Thanks to the D. Fitzpatrick, A. Jackson, and M. Dunlop clinical groups within the IGMM, for the data they have shared.

Thanks to the Caceres lab for their collaboration on the microRNA processing project.

Contents

List of Figures

List of Tables

Abbreviation	Description
CAGE	cap analysis of gene expression
ceRNA	competitive endogenous RNA
ChIP	chromatin immunoprecipitation
CLASH	cross-linking ligation and sequencing of hybrids
CLIP	cross linking and immunoprecipitation
DSB	double strand break
endo-siRNA	endogenous siRNA
eQTL	expression quantitative trait loci
eRNA	enhancer RNA
ESCs	embryonic stem cells
ExAC	exome aggregation consortium
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
lincRNA	long intergenic non-coding RNA
lncRNA	long non-coding RNA
MAF	minor allele frequency
MOPD	microcephalic osteodysplastic primordial dwarfism
mRNA	messenger RNA
OR	odds ratio
piRNA	PIWI-associated RNA
pre-microRNA	precursor microRNA
pri-microRNA	primary microRNA
PTGS	post-transcriptional gene silencing
RISC	RNA-induced silencing complex
RNAi	RNA interference
rRNA	ribosomal RNA
sdRNA	double stranded RNA
SHAPE	Selective 2‘-hydroxyl acylation analysed by primer extension
shRNA	small hairpin RNA
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
snoRNP	small ribonucleoprotein complex
snRNA	small nuclear RNA
TPM	transcripts per million
tRNA	transfer RNA
TSS	transcription start site

Table 1: Abbreviations

Chapter 1

Introduction

Small RNAs are typically ∼20-30nt in length identified in animals, plants and fungi with a variety of known roles in gene regulation. They can be categorised into several broad classes of small RNA species including microRNAs, PIWI-associated RNAs (piRNAs) and small interfering RNAs (siRNAs) with others continually being added with varying production and effector pathways. Small RNAs act through interactions — inter- or intra-molecular base pairing RNA–RNA interactions where A pairs with U and G with C, and RNA–protein interactions — in the formation of functional complexes, best known as regulators of gene expression via base-pairing to messenger RNA (mRNA) affecting the production of their protein products.

1.1 Discovery of small RNAs

1.1.1 microRNAs

Small RNA directed gene regulation was discovered in the nematode worm *C. elegans* where the gene *lin-4* was identified through a developmental screen as a negative trans regulator of *lin-14*. *Lin-4* controls the transition from the first to the second larval stage. It was located through positional cloning and found to be unlikely to encode a protein. Two transcripts were found corresponding to the *lin-4* locus, one 61nt in length and the other 22nt — which would now be called the pre-microRNA and microRNA respectively — this smaller RNA being complementary to the 3'UTR of the *lin-14* gene [1, 2]. Lee et al observed that "*lin-4* may represent a class of developmental regulatory genes that encode small antisense RNA products".

A second regulatory small RNA was discovered again through a developmental screen in *C. elegans*, this time for genes which suppressed the synthetic sterile phenotype of a strain with *lin-14* and *egl-35* mutants [3]. Mutants of *let-7* were strongest to retard the heterochronic defects of the *lin-14* background, and mutants of *let-7* displayed lethal phenotypes with larval

cell fates being reproduced during the adult stage of development. Similarly to *lin-4*, *let-7* was found to correspond to no conventional (protein-coding) gene product, however evolutionary conservation analysis between *Caenorhabditis* species identified a conserved 26-bp region overlapping one *let-7* point mutation. Northern blot analysis detected a 21nt RNA transcript from this conserved region. Expression of this transcript was found to be temporally regulated – expressed only in the L4 and adult stages – in agreement with its observed functions. Regions complementary to *let-7* were found in the experimentally determined sequences of heterochronic genes, present only in the 3'UTRs of *lin-14, lin-28, lin-41, lin-42* and *daf-12*. $\sim$21nt small RNAs such as *lin-4* and *let-7* came to be known as microRNAs.

As more genome sequences were becoming available evolutionary analysis identified similar small RNAs in a variety of other species. Initial studies found homologs of *lin-4* only in the genus *Caenhorabditis*, however homologs of *let-7* were found in a wide range of animal species [4].

The discovery of additional microRNAs has proceeded through three main approaches: capture, isolation and sequencing; forward genetics; and computational prediction. Other than the initially discovered *lin-4* and *let-7* few microRNAs have been identified in loss of function genetic screens. Through comparative genomics and cDNA cloning other microRNAs were identified in *C. elegans* some of which were temporally regulated during development with potential orthologs in *Drosophila* and humans [5, 6, 7]. Direct cloning or sequencing of small RNAs[8, 9, 10] and computational prediction have been the most commonly used methods to detect microRNAs. Computational methods relying on the prediction of hairpin structures where RNA molecules fold and base-pair with another section of the same strand forming a U shape, and evolutionary conservation of microRNA sequences[11] where some microRNAs being highly conserved while others are evolving rapidly[12].

Other small RNA types have also been classified including siRNAs and piRNAs and snoRNAs which share some sequence, structural or functional properties of microRNAs but are sufficiently different to warrant a distinct classification.

1.1.2 siRNAs

The introduction of small double stranded RNA (dsRNA) molecules into cells was found to interfere with endogenous RNAs more potently than either sense or antisense RNA individually. These dsRNAs are processed to become small interfering RNAs (siRNAs), small RNA molecules similar to microRNAs, and associate with the same protein complex to silence specific targets via complementary base pairing, functioning in human cell culture[13]. This effect was termed RNA interference (RNAi)[14], and its discoverers Andrew Fire and Craig Mello were awarded the 2006 Nobel Prize in physiology or medicine. Since then small hairpin RNA (shRNA) libraries of dsRNAs have been developed as a tool for loss-of-function screens[15] and are now widely used. Also RNAi based therapeutics have been under

investigation with some advantages over traditional small molecule drugs such as specificity, however delivery strategies that can be used in clinical settings remain a major challenge [16].

1.1.3 piRNAs

PIWI-interacting RNAs (piRNAs) are 21–30nt single stranded RNAs named for their association with PIWI proteins. PIWI proteins were identified as necessary for germline integrity: *piwi* was found to be necessary for self-renewal of germ cells in *Drosophila*[17]. *Aubergine*, another PIWI clade member is necessary for pole cell formation to produce functional oocytes [18]. Similar roles were found in mammals with *miwi* (murine *piwi*) being essential for spermatogenesis [19] and *Mili* (*Miwi* like *Piwil2* (*piwi* like 2)) also being essential for spermatogenesis in mice [20]. Mutations in members of the PIWI gene family were also seen to affect mobile genetic elements, with *piwi* silencing the endogenous retrovirus *gypsy*, and *aubergine* repressing the telomere retroelement *TART*. Identification of the guide RNAs associated with these PIWI clade proteins initially identified RNAs 29-30nt in length in mice and rats which were mostly clustered into genomic regions less than 100kb in length with expression in the testes essential to spermatogenesis [21, 22, 23, 24].

1.1.4 snoRNAs

Small nucleolar RNAs (snoRNAs) are a class of non protein-coding small RNAs longer than the previous examples at ∼70-200nt, best known for guiding the modification of other non protein-coding RNAs including ribosomal, small nuclear and transfer RNAs (rRNA, snRNA and tRNA respectively). U3 - the most abundant snoRNA has long been known to bind to pre-ribosomal RNA[25], with snoRNAs functioning through the direct base-pairing of part of their sequence[26].

1.2 Argonaute proteins

Three types of small RNAs discussed here (microRNA, siRNA and piRNA) act primarily through binding to members of the argonaute family of proteins. These were first observed in the *C. elegans* RNA interference deficient (rde) mutants [27], which were discovered in a screen for RNAi mutants by mutagenising wild-type animals and culturing them on a bacterial lawn expressing dsRNA complementary to the essential gene *pos-1*. The F2 generation could then be searched for individuals able to produce viable progeny. These mutant strains were subsequently examined for their RNAi activity in somatic cells by injecting dsRNA targeting the collagen gene *sqt-3* and the body muscle structural gene *unc-22* and examining animal shape and the presence of the body twitching phenotype. The *rde-1, rde-3, rde-4 and mut-2* activities appeared to be required for RNAi of all genes analysed whereas *rde-2* and *mut-7*

activities were specific to RNAi in the germline but not somatically expressed genes [27]. This observation led to the functional distinction between production of siRNA from longer dsRNA, and the subsequent utilisation of the produced siRNAs [28]. *rde-1* and *rde-4* were seen to form a complex with dicer (*dcr-1*) an RNAseIII nuclease responsible for cleaving dsRNAs in the production of small RNAs for RNAi [29, 30].

The argonaute family of proteins in animals can be divided into two main subfamilies, the AGO subfamily or clade are named for their similarity to the *Arabidopsis thaliana* AGO1 protein, and the PIWI clade which are homologous to the *Drosophila melanogaster* Piwi protein [31]. AGO proteins bind microRNAs and siRNAs, leading to post-transcriptional gene silencing (PTGS or RNAi). Whereas PIWI proteins bind piRNAs and are expressed mainly in germ cells where they repress transposable elements.

Humans have eight argonaute family members, Piwi-like 1-4 (PIWIL1, PIWIL2, PIWIL3, PIWIL4) of the Piwi clade and human AGO 1-4 (AGO1, AGO2, AGO3, AGO4) of the AGO clade [31]. The AGO members are closely related and appear to be ubiquitously expressed. While in some species different AGO associated small RNA species appear to bind to distinct AGO proteins, human AGO1 through AGO4 seem to bind indiscriminately to microRNAs [32]. Functionally human AGO proteins appear to be equivalent. ES cells deficient for all four AGO proteins lose microRNA silencing and undergoing apoptosis, with rescue from reintroduction of any single AGO [33]. However only AGO2 is catalytically active and able to cleave bound RNA molecules [34, 35] and there is evidence for some microRNAs and other small RNAs being preferentially associated with distinct AGO proteins [36].

1.3 Small RNA processing

1.3.1 microRNAs

microRNAs are encoded within the genome as larger transcripts known as primary microRNAs (pri-microRNAs), where they form hairpin loops in the RNA secondary structure (Fig 1.1a). These structures are then processed in a series of steps first in the nucleus and then in the cytoplasm to produce mature microRNAs incorporated into the RISC (Fig 1.1b).

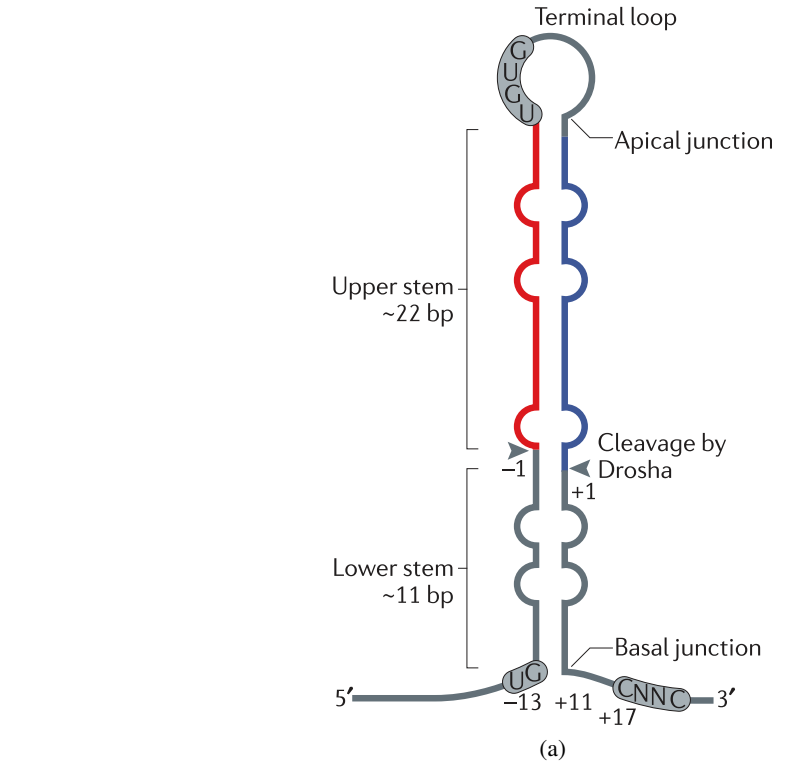
Primary microRNAs (pri-microRNAs) are processed into precursor microRNAs (pre-microRNAs) by the nuclear microprocessor complex containing Drosha and DGCR8 [37, 38]. Conserved in mammals and *C. elegans* Drosha is an RNase III enzyme capable of cleaving pri-microRNA to release pre-microRNA *in vitro* [39]. DGCR8 (DiGeorge critical region 8, also known as Pasha – partner of Drosha) is a double stranded RNA binding protein which positions the Drosha cut site 11bp from the base of the hairpin stem, to release the $\sim$70nt pre-microRNA[40].

While recognition of pre-microRNA hairpins by Drosha/DGCR8 remains to be fully

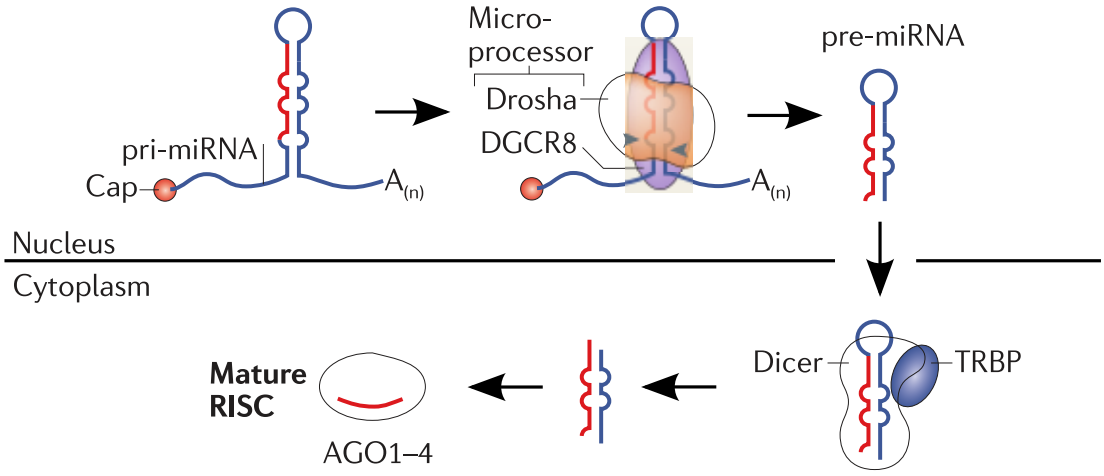
elucidated, with many predicted microRNAs failing to produce mature microRNAs *in vivo* and some loci with non-canonical stem base pairing patterns producing mature microRNAs[41], key sequence and structural motifs have been identified: 40nt flanking either side of the miR-223 pre-microRNA were seen to be necessary for processing *in vitro*[42]. Including an $\sim$11nt stem – or one turn of the RNA helix – flanking the pre-microRNA which determines the cleavage site, and the single-stranded RNA regions flanking this are also critical for processing[43].

Analysis of many variants of four human pri-microRNAs identified three sequence motifs seen in Fig 1.1a: a basal UG present in 24% of human microRNAs, a UGUG motif in the terminal loop present in 20% of human microRNAs and a 3' CNNC motif present in 30% of human microRNAs[45]. This CNNC motif was associated with SRp20/SRSF3 binding, a splicing factor with an RNA recognition motif. Analysis of additional microRNA variants in a high throughput assay revealed a preference for base-pairing in microRNA stems in all but one position, although fully or near fully paired microRNA stems may have additional cellular consequences in producing an interferon response, this study also confirmed the positive effect on processing of the basal UG and apical UGUG motifs[46]. However $\sim$20% of human microRNAs lacked all of these sequence motifs suggesting additional factors or recognition features remain to be discovered. The RNA modification N6-methyladeosine (m6A) has also been suggested as a mark recognised by DGCR8 to promote processing, deposited by the methyltransferase-like 3 (METTL3) protein to GGAC motifs in the pri-microRNA[47]. It should also be noted that all features defined to date relate to primary sequence or predicted secondary structure, whereas processing is likely to occur in the context of three-dimensional folding for which we currently have few measures or estimates.

Pre-microRNAs are exported to the cytoplasm via Exportin-5, a RanGTP dependent RNA binding protein which traffics pre-microRNA through the nuclear envelope as an exportin - RanGTP - cargo complex where the cargo and Ran are released in the cytoplasm after GTP hydrolysis [48, 49]. In the cytoplasm pre-microRNAs are further cleaved by Dicer to remove the hairpin loop. Knockdown of dicer in human cells leads to the accumulation of pre-microRNA [50]. Dicer cleaves pre-microRNAs $\sim$22nt from the base of the ds-RNA stem close to the terminal. Cleaved microRNAs are transferred to an AGO protein to form the RNA-induced silencing complex (RISC) – the protein complex which represses target gene expression[51]. The RISC is further mediated by several other proteins: In humans TRBP (transactivation-response (TAR) RNA binding protein) a ds-RNA binding protein functions as an asymmetry sensor [52] and the HSP70/HSP90 chaperone machinery is required to load small RNA duplexes into argonaute proteins [53, 54]. microRNA duplexes are generally classified as having a guide and a passenger strand. The guide strand is predominantly the active strand and is incorporated into argonaute while the passenger strand is degraded [55, 51], however passenger strands have also been seen to be incorporated into the RISC sometimes in a tissue specific manner [56].



(a)



(b)

Figure 1.1: (a) Annotated microRNA stem loop showing sequence motifs and RNA structural features commonly found.
(b) - microRNA processing schematic showing nuclear processing by the microprocessor, cytoplasmic processing by Dicer, and incorporation into Argonaute to form the mature RISC. Adapted by permission from Macmillan Publishers Ltd: [Nature Reviews Molecular Cell Biology] [44], copyright (2014)

Pre-microRNAs may also be generated in a Drosha independent pathway as short introns (Mirtrons) which are excised by the spliceosome [57, 58] and then exported to the nucleus for further processing.

Another source of microRNAs are the transcriptional start site associated microRNAs (TSS-microRNAs) observed in small RNA sequencing data[59] and immunoprecipitated AGO protein complexes where enriched promoter proximal regions with bimodal peaks corresponding to the 3p and 5p arms of a Dicer-processed product[60]. These TSS-microRNAs were found at RNApolII paused regions and contained more structured transcripts compared to control TSSs, though the functional significance of these TSS associated, dicer processed transcripts remains unclear[61].

1.3.2 piRNAs

Single stranded RNA molecules are the precursors to piRNAs, as such they are Dicer independent, transcribed in most species from clusters of piRNA genes producing precursor transcripts from a few kb to 100kb in length[23]. However in *C. elegans* each piRNA is encoded as a small transcriptional unit. Two main pathways for the biogenesis of piRNAs have been observed. In the primary pathway precursor piRNAs transcribed from the genome undergo processing at specific sites at the nuclear pores of germ cell nuclei where they are cleaved by a nuclease enzyme to generate the 5' end of the small RNA, this nuclease has been suggested to be the Zucchini protein which in *Drosophila* mutants display primary biogenesis defects [62]. After generation of the 5' end piRNAs are loaded onto PIWI proteins requiring chaperone proteins, notably HSP90 or its orthologs [63]. piRNA lengths have a broad distribution, the RNAs associated with different PIWI proteins having different length distributions. The enzymes responsible for 3' trimming to produce the final 3' end are yet to be determined but are followed by 2'OH methylation of the 3' base protecting the piRNA from uridylation or destabilisation [64, 65]. In the secondary piRNA pathway (sometimes referred to as ping-pong amplification) cleavage of targets by piRNA generate the 5' end of a new piRNA allowing an adaptive response to target expression [66, 67].

piRNAs generally target transposable elements [68] during germline development where they are expressed, although other target sequences have been observed including developmental genes in *Drosophila* [69] and genes involved in the establishment of long term memory in sea slug neurones[70]. PIWI proteins may silence their targets via RNA degradation [66, 67], transcriptionally or post-transcriptionally. Transcriptional regulation has been suggested as some PIWI homologs are observed in the nucleus after loading a piRNA and the PIWI homologs MIWI and MILI2 are also localised to the nucleus where they lead to DNA methylation of target loci [71, 72].

1.3.3 Endogenous siRNAs

Another category of small RNA derived in a Dicer independent pathway are the endogenous siRNAs (endo-siRNAs). These are produced by RNA-dependent RNA polymerases (RdRPs) from cellular mRNAs, observed in *C. elegans* [73, 74] *Drosophila* [75, 76, 77] and mice [78, 79]. Derived from: transposable elements [78, 79]; cis-natural antisense transcripts (cis-NAT) [79, 75, 77]; and trans-NATs derived from gene-pseudogene pairs [78, 79].

1.3.4 snoRNAs

In vertebrates snoRNAs are generally excised from the introns of pre-mRNAs although some have their own promoter, and are generated through cleavage by exo- and endonucleases. Divided into two main classes: C/D box snoRNAs bind the proteins of the small ribonucleoprotein complex (snoRNP) including Fibrillarin, a methyltransferase that catalyses the 2'-O-methylation of ribose in target RNAs. As with siRNA, microRNA and piRNA targeting occurs through complementary base pairing. The snoRNAs contain conserved complementarity to the universal core regions of rRNAs. H/ACA box snoRNAs bind proteins to form snoRNPs which function in pseudouridylation of target sequences [80, 81]. Some microRNA-like small RNAs can originate from snoRNAs, requiring Dicer for processing and associating with AGO1 and AGO2 [82].

1.4 Targeting by Complementarity

After microRNAs were discovered as a large class of regulatory elements identifying their targets became an important task. In plants this is less complex as extensive complementarity can be used to predict interactions [83]. However in animals complementarity between a region of the target and the whole length of the microRNA is not necessary to target the RISC and algorithms to predict microRNA target sites were developed [84, 85, 86].

1.4.1 microRNA – Target evolution

Analysis of microRNAs and their targeting suggests that although some microRNA:target relationships such as *let-7:lin-41* persist over long evolutionary distances[4], overall microRNA-target relationships appear to be conserved close to the level expected by chance, with more rapid turnover than other regulatory mechanisms during evolution [87, 88]. Within humans microRNA target sites appear to be under stronger negative selection than other conserved sequence motifs in 3'UTRs[89].

One study identifying conserved microRNA response elements in human protein coding 3'UTRs found more than 45000 microRNA response elements in greater than 60% of

genes[90].

A search for novel human microRNA genes uncovered miR-941 which evolved after the separation of the human and chimpanzee lineages with variable precursor copy number in the human genome. miR-941 was found to be expressed in a variety of tissues at high levels including the brain, with targets suggesting roles in cellular differentiation. This was associated with an accelerated loss of miR-941 binding sites in the human genome, presumably to escape gene regulation by this newly evolved microRNA[91].

Collectively these results suggest that sequence conservation or its absence are of little predictive value for microRNA identification or inferring microRNA-target interactions.

1.4.2 Computational Prediction

The algorithms initially developed to identify interactions utilised a number of features when scoring a microRNA target sequence:

- Restricting searches to the 3'UTR of protein coding genes to reduce the search space[84].
- Watson-Crick pairing in the 5' region of the microRNA around nucleotides 2-7 — the seed region — were considered most important and searching for 7nt matches to the seed region was generally the first step[92].
- Orthologous 3'UTRs compiled from whole genome alignments could then be used to examine the evolutionary conservation of the Watson-Crick pairing of the seed match[85].
- Regions of base-pairing in addition to the seed regions can be analysed to supplement and strengthen the prediction or to compensate for lower seed pairing or conservation.

Changing parameters controlling the stringency of seed match necessary and the requirement for evolutionary conservation or additional pairing allow for altering the sensitivity and specificity of results. These predictions concluded that some — highly conserved — microRNAs have a high number of conserved targets suggesting that more than half of human protein coding genes appear to maintain microRNA target sites in their 3'UTRs[92]. Introduction of microRNAs into cell culture demonstrated modest effects on hundreds of genes[93].

The incorporation of results from high throughput direct assays of microRNA-target interactions has improved the target prediction algorithms[94, 95]. Analysis of CLASH derived microRNA - mRNA pairs has been used in the development of TargetScan v7[96] which is considered the current state-of-the-art in microRNA target prediction.

1.4.3 Direct Assay

Methods to probe these small RNA interactions have benefited from the increasing availability of DNA sequencing in order to probe these interactions genome-wide. CLIP (Cross Linking and Immunoprecipitation) uses UV to crosslink proteins with RNA, followed by immunoprecipitation for proteins of interest and cDNA sequencing. Chi et al performed HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking and immunoprecipitation) to examine AGO targeting in the mouse brain generating interaction maps for the 20 most abundant microRNAs with the most abundant being miR-124, they also found 27% of AGO-mRNA clusters had no predicted seed match among the top 20 most abundant AGO-microRNA families [97]. Hafner et al used PAR-CLIP (photo-activatable ribonucleoside enhanced crosslinking and immunoprecipitation) — a modification of CLIP which incorporates photo-activatable nucleoside analogues into RNA and significantly improved RNA recovery — to investigate a number of RNA binding proteins including AGO1-4 from HEK293 cells. They found that 84% of binding sites were present within exonic regions – of those 4% to the 5'UTR, 50% to the coding sequence and 46% to the 3'UTR. Transfection of cells with a cocktail of oligonucleotides to block the 25 most abundant microRNAs was performed to examine the function of these targets. Genes with target sites exclusively within their coding sequence had negligible effect compared to the upregulation of genes with targets sites within their 3'UTR. The length of seed match had a noticeable effect on transcript abundance as targets with 9nt seed matches showing the greatest upregulation after microRNA inhibition [98]. Leung et al performed CLIP for AGO2 in mouse embryonic stem cells (mESCs) finding 201 enriched motifs in 3'UTRs and 103 within coding sequence. They also found that different to other cell types analysed a microRNA cluster – the miR-290-295 cluster appeared to be responsible for the majority of targets [99].

Kudla et al observed during analysis of cross-linking and analysis of cDNAs (CRAC) of the yeast RNA helicase Prp43 that some reads contained chimeric cDNAs with the guide region of a snoRNA fused to a target site in the 18S rRNA, they hypothesised that during the ligation of linker sequences in the CRAC procedure RNA molecules could be ligated together. Computationally analysing the CRAC reads to identify these chimeric examples found 0.46% of reads composed of distinct fragments which could be mapped to different RNA molecules or different regions of the same molecule, this analysis was named cross-linking ligation and Sequencing of Hybrids (CLASH) [100]. Application of the CLASH technique to microRNA interactions (Fig 2.1) was performed via UV cross-linking and ligation, followed by purification of PTH-AGO1 (N-terminal fusion of hAGO1) using six different experimental conditions, and results from all six protocols were analysed together [101]. They found $\sim$2% of the total reads were chimeric with 70% of microRNA targets present within mRNAs, of these 5.5% were within 5'UTRs, 33.5% were within 3'UTRs and 61% were within coding sequence, similar to results seen by CLIP methods. They also found a large class of chimeras with base pairing only present at the 3' end of the microRNA representing 18% of microRNA-mRNA

interactions suggesting that previous CLIP and computational prediction of microRNA target methods may have been biased towards finding interactions with Watson-Crick base pairing within the seed region.

1.5 Mechanisms of regulation by microRNAs

1.5.1 Post-transcriptional regulation

The regulation of target mRNAs by microRNAs in animals are generally thought to occur at the level of translation, whereas in plants they were though to act through target cleavage due to the more extensive base pairing between microRNAs and their target sites in plants [1, 83, 102].

The mechanism of mRNA silencing by microRNAs occurs through a combination of translational repression and mRNA destabilisation via decapping, deadenylation and 5' to 3' mRNA degradation, the relative contributions of these processes remains a research question with evidence that both mRNA destabilisation[103, 104] or translational inhibition[105, 106] cause the majority of this effect, although the processes may be linked[107].

Repression is mediated through the interactions of Argonaute, which forms the core of a complex of proteins, including RNA helicases and mRNA binding proteins[108]. GW182 proteins, required for silencing, are a key component of this complex of proteins forming interactions with cytoplasmic deadenylase complexes[109].

Mechanisms of translational repression by Argonaute have been reported to involve interactions with the eIF4A RNA helicase proteins, which unwind mRNAs allowing the 43S pre-initiation complexes to scan for the presence of the start codon[105, 106].

1.5.2 Transcriptional Gene regulation

It has now become clear that shortRNA directed gene regulation occurs at the levels of translational repression and target degradation in both plants and animals[110, 103, 42, 111]. Cleavage of target mRNAs occurs through perfect complementarity between microRNA and target, leading to cleavage by a catalytically active AGO protein [102]. Translational repression of microRNA target sequences can occur at several stages during translation: The initiation of translation can be blocked, seen though experiments using cell free extracts to compare translation of native mRNAs to those with internal ribosome entry sites (IRES) [112, 113]. Post initiation mechanisms were suggested by the observation that microRNAs and their targets were associated with polysomes in sucrose sedimentation gradients [114, 115]. A number of different mechanisms have been suggested to contribute to post-initiation translational repression including inhibition of elongation, co-translational degradation and premature termination of translation [116].

Increasing amounts of sequencing data have also increased our understanding of transcriptional complexity. Sense-antisense transcripts have been observed with roles in gene expression in many organisms including the production of endo-siRNAs [117, 118].

Nuclear functions for argonaute have been well observed in fungi and plants, and have been suggested but are under explored in mammals[119, 120]. In the fission yeast *Schizosaccharomyces pombe* deletion of the RNAi machinery resulted in the accumulation of transcripts from centromeric heterochromatic repeats [121], and loci located within heterochromatin regions displaying bidirectional transcription co-localised with Dicer1 at the nuclear periphery favouring the production of endo-siRNAs [122].

In plants RNA has been shown to direct DNA methylation at specific, complementary sequences. This was first observed using tobacco plants engineered to carry viroid identical DNA sequences within their genomes which became methylated in strains carrying the actively replicating viroid [123].

In the ciliate *Tetrahymena* a member of the PIWI subfamily is required for programmed DNA elimination. This occurs in ciliates due to the presence of the germline micronucleus and somatic macronucleus, during fertilisation DNA rearrangements and sequence specific DNA elimination occur to leave the macronucleus lacking $\sim$15% of the DNA sequences present in the zygotic nucleus or micronucleus. TWI1 (*Tetrahymena* PIWI1) is required for the formation of viable progeny and knockout of TWI1 leads to loss of internal eliminated sequence (IES) excision related to H3K9 methylation [124, 125].

The *Drosophila* PIWI homolog has been observed to have functions in the nucleus, associating with heterochromatin protein 1 alpha (HP1a) [126] and loss of *Drosophila* PIWI leading to a reduction of H3K9me3 at targeted loci [127, 128].

AGO proteins have also been associated with nuclear functions in *Drosophila* where AGO2 and Dicer2 have been seen to associate with transcriptionally active loci and the core transcription machinery. After heat shock null mutants for AGO2 and Dicer2 impair the global dynamics of RNAPolII [129]. Moshkovich et al used AGO2 ChIP-seq finding it to be localised to euchromatin and in particular co-localising with CTCF/CP190 chromatin insulators [130].

AGO2 is predominantly located within cytoplasmic processing bodies [131], however AGO proteins as well as mature microRNAs have been observed in the nucleus via immunoflourescence and cellular fractionation techniques [132, 133, 131, 134, 10, 135, 136].

1.5.3 shortRNA directed transcriptional gene regulation in mammals

Nuclear functions of AGO in transcriptional gene regulation have also been observed in mammalian cells but remain controversial with some studies suggesting a repressive role, and others suggesting an activating role for AGO in the nucleus.

Suggesting a repressive role siRNAs targeted to the promoter region of an EF1A promoter – GFP reporter gene as well as endogenous EF1A finding silencing of both the reporter and endogenous gene which was associated with DNA methylation of the targeted sequence [137]. Short hairpin RNAs (shRNAs) complementary to the RASSF1A promoter were observed to direct low levels of DNA methylation and partial gene silencing in HeLa cells[138]. Transfection of HCT116 human colorectal cancer cell lines with double-stranded oligonucleotides homologous to the CpG island of the CDH1 promoter found that CDH1 protein levels were repressed without a change in the level of DNA methylation [139]. siRNAs induced a potent knockdown of 7SK small nuclear RNA (snRNA) in the nucleus [132]. siRNAs targeting upstream of the transcription start site at the binding sites for transcription factors reduce expression levels seen by nuclear run-on assay, this effect seemed to be dependant on both AGO1 and AGO2 [140]. miR-320 encoded antisense within the promoter region of POLR3D was able to direct AGO1, polycomb group component EZH2 and H3K27me3 to the POLR3D promoter suggesting a cis-regulatory role at this locus directing transcriptional gene silencing [141]. microRNA mimics, consisting of microRNA the sequence and a fully complementary RNA carrier strand, predicted to target the progesterone receptor gene promoter inhibit gene expression associated with decreased RNAPolII occupancy and increased H3K9 dimethylation [142]. Promoter profiling with ChIP-on-chip using an antibody targeting all human AGO proteins in senescent and presenescent WI38 fibroblasts, found AGO2 associated with the retinoblastoma (RB1)E2F repressor complex [143]. And suggesting an activating role for AGO in the nucleus dsRNA targeting several gene promoters led to long-lasting sequence-specific induction of the target genes, associated with a loss of H3K9 methylation [144], and chromatin immunoprecipitation and sequencing (ChIP-seq) performed on HA-AGO1 and HA-AGO2 expressed in PC3 – a human prostate cancer cell line to examine the distribution of AGO proteins in the nucleus. Finding that AGO1 but not AGO2 was associated with the promoters of actively transcribed genes [145]. Cell fractionation studies have shown that RNAi factors and microRNAs are located in the nucleus and that siRNAs targeted to nuclear lncRNAs were able to repress their targets[146]. A SILAC study demonstrated TSS proximal small RNAs loaded into AGO2 associated with the SWI/SNF complex[147]. AGO2 was seen to associate with the H3K9 methyltransferase SETDB1 suggesting a mechanism of chromatin remodelling in TGS[148]. The formation of R-loops, where DNA and RNA strands base pair, form at RNAPolII paused sites may also be a source of dsRNAs processed by DICER and loaded onto AGO proteins[149].

1.5.4 Argonaute and small RNAs in splicing

Several studies have also implicated AGO proteins in affecting alternative splicing.

siRNAs targeted near an alternative exon were found to affect the alternative splicing at that exon in HeLa and hepatoma cells, this effect being dependant on AGO1 and suggested to be due to affecting the elongation rate of RNAPolII as the heterochromatin marks H3K9 dimethylation

and H3K27 trimethylation were deposited at the target and the effect was reduced or abolished in treatments which promote chromatin relaxation or an increased RNAPolII elongation rate [150].

In mice AGO1 and AGO2 coimmunoprecipitated with the spliceosomal small nuclear RNP (snRNP) U2 and U5 subunits, and using the CD44 gene as a model for alternative splicing as it encodes a cluster of nine variant exons which display increased inclusion upon stimulation of protein kinase C by the compound PMA, depletion of AGO1 or AGO2 compromised this exon variant inclusion, reducing the recruitment of H3K9 trimethylation and slowing RNAPolII progression [151]. In *Drosophila* Ago2 ChIP-seq found most peaks at gene promoters as well as Ago-2-null mutants showing defects in pre-mRNA splicing patterns [152].

An AGO1 ChIP experiment demonstrated that in the nucleus AGO1 binds primarily at enhancer regions, mediated by enhancer RNAs (eRNAs)[153]. However depletion of AGO1 led to changes in constitutive and alternative splicing rather than transcription, this is suggested to occur through changes in the chromatin marks present at gene loci, which in turn affect the rate of RNAPolII elongation[153].

Models of the differences in splicing between MCF7 an MCF10 which including binding sites of CTCF, AGO1, HP1 and chromatin marks were able to explain $\sim$69% of splicing changes between the two cell types, with AGO1 binding clusters associated with CTCF and HP1 binding sites[154].

1.5.5 Double-strand break associated small RNAs

Small RNAs are produced from regions of double strand breaks (DSBs) in human U2OS (osteosarcoma) cells and *Arabidopsis* where they associate with Dicer or Dicer-like proteins and AGO2, and knockdown of Dicer or AGO2 impairs the efficiency of DSB repair [155, 156]. AGO2 has also been seen to interact with Rad51, an interaction which is enhanced in the presence of ionising radiation, with AGO2 suggested to promote the recruitment of Rad51 at DSBs[157].

1.6 microRNA sequestration by competitive endogenous RNA (ceRNA)

It has been hypothesised that competition between transcripts with shared microRNA binding sites can affect the post transcriptional regulation of those microRNAs, with competing transcripts including long non-coding RNAs (lncRNAs), pseudogenes and circular RNAs.

Large numbers of circular RNAs have been postulated or identified in human cells [158, 159, 160] with suggested roles acting as microRNA sponges due to observed high numbers of microRNA binding sites, one example the cerebellar degeneration-related protein 1 transcript

(CDR1as) containing 63 conserved binding sites for miR-7 and is densely bound by AGO in PAR-CLIP datasets [160]. Circular RNA products from chromosomal translocations in cancer have also been suggested to contribute to cellular transformation[161].

However the vast majority of circular RNAs did not demonstrate any selective constraint above flanking exons or contain more microRNA target sites than would be expected by chance, CDR1as being only one of two predicted circular RNAs with more microRNA target sites than expected[162]. This result suggested that the vast majority of circular RNAs represent low abundance alternatively spliced isoforms with debatable biological significance.

Several theoretical and practical studies have questioned whether ceRNAs will be likely to have enough microRNA target sites, or be expressed at sufficient levels to influence target repression by microRNAs: Altering the abundance of a validated miR-122 target in liver cells suggested that a ceRNA must approach the normal target site abundance ($1.5x10^5$ for miR-122 in liver cells) to affect target repression in a detectable manner[163]. Studies using predicted target sites and reporter assays suggested that microRNA susceptibility to ceRNAs depends on the relative microRNA:target ratio, with only those microRNAs with low microRNA:target ratios being susceptible to ceRNA effects under normal circumstances[164, 165, 166, 167].

Additional examples of ceRNAs have been reported including: Hmga2 a non-histone chromosomal protein with seven let-7 target sites in its 3'UTR, this activity important in progression of non-small-cell lung cancer[168]. The BRAF pseudogene, acting as a ceRNA to affect the expression of BRAF, which can act as a proto-oncogene[169]. Due to the intrinsic sequence similarity of this gene/pseudogene pair these transcripts share many high affinity microRNA target sites increasing the likelihood of ceRNA crosstalk. And Lin28B acting as a ceRNA for let-7 in neuroblastoma cell lines leading to de-repression of MYCN, amplification of which is associated with poor prognosis in neuroblastoma[170].

lncRNAs are an additional potential source of ceRNAs, one lncRNA knockdown study suggesting that one fifth of transcript level changes induced by lncRNA knockdown were due to microRNA mediated crosstalk[171]. Several other lncRNA microRNA sponges were predicted computationally using AGO CLIP data[172].

1.7 microRNA polymorphisms

Polymorphisms at microRNA loci have the potential to impact their function in a number of ways: Disruption of protein binding sites affecting microRNA processing. Disruption of the RNA structure through changing base-pairing in the stem-loop, leading to less efficient processing. Changes in the mature microRNAs may affect their incorporation into the argonaute protein, or could change the targets recognised by the base-pairing of the microRNA - particularly if the polymorphism is in the microRNA seed region.

1.7.1 microRNA polymorphisms in the human population

Catalogues of polymorphisms in microRNAs and microRNA target sites have been generated over the past decade:

- In 173 pre-microRNA regions in a cohort of 96 samples representing the general population in Japan finding 10 SNPs in 10 pre-microRNAs[173].
- In 227 pre-microRNA regions in dbSNP finding 323 SNPs, 12 within the pre-microRNAs including a SNP within the miR-125a mature microRNA which affected the processing from pri-microRNA to pre-microRNA[174].
- In 474 pre-microRNAs in dbSNP and also in predicted target sites, finding 65 SNPs in 49 pre-microRNAs, a lower SNP density ($\sim$1.3 SNPs per kb) than flanking – often intergenic – regions ($\sim$3 SNPs per kb)[175].
- 15 SNPs were found within predicted microRNA target sites for 125 cancer associated genes which altered the binding energy for the predicted interaction[176].
- Whole-genome sequences from 1092 individuals sequenced in the 1000 genomes project and 60 exome sequences from healthy individuals in the south of Spain were used to identify 527 SNPs including 45 within the microRNA seed regions[177].

Several databases have automated the collection and annotation of common SNPs in pre-microRNAs and in predicted microRNA target sites[178, 179, 180, 181].

1.7.2 microRNA eQTLs

Using expression arrays or short RNA sequencing in populations of genotypically diverse cells, associations between SNPs and the expression of microRNAs can be quantified. This is done by correlating the SNP genotypes with gene expression patterns in order to identify expression quantitative trait loci (eQTLs). These SNPs represent those which may affect the function of microRNAs through altering their expression level - which could occur though affecting the transcription, nuclear processing, cytoplasmic processing or stability of microRNAs. cis- or local eQTLs have been defined as those within 50kb to 1MB of the target gene depending on the study, with other associated SNPs being trans-eQTLs. It is expected some, but not all, cis-eQTLs will be due to altered microRNA processing as variants can effect sequence or structural recognition features.

- 180 primary fibroblasts were genotyped on Illumina Hap550 SNP array and had expression profiling on the TaqMan microRNA Array V1 finding cis-eQTLs (within 1Mb) for 12 microRNAs[182].
- 70 samples of abdominal adipose tissue were genotyped on the Illumina BeadChip 317k and microRNA expression profiled with Illumina microRNA BeadArray finding

6 independent cis-microRNA eQTLs[183].

- small RNA sequencing and genotyping of 131 samples of adipose tissue were used to find 14 cis-eQTL loci, 7 of which were also eQTLs for a mRNA transcript in the region[184].
- microRNA expression was quantified with the Exiqon miRCURY array in 60 European and 60 African lymphoblastoid cell lines (LCLs) sequenced as part of the International HapMap Project, finding 31 genome-wide significant SNP associations with microRNA expression[185].
- small RNA sequencing in 363 European and 89 African individuals sequenced as part of the 1000 genomes project found 60 microRNAs with cis-eQTLs[186].

Few studies have gone on to the next step to investigate the mechanisms by which identified microRNA eQTLs effect expression level and none have done-so systematically for multiple microRNAs or variants. Those which have been investigated are mainly SNPs or rare variants identified as associated with diseases, particularly cancers, and will be discussed in the next section.

1.7.3 microRNA polymorphisms in disease

microRNA variants may cause disease through their effects on microRNA function as outlined above: altering transcription of the microRNA precursor, the biogenesis of the mature microRNA from the precursor or by altering microRNA-target interactions. Case-control studies have found evidence for associations between common microRNA variants and diseases, particularly in cancer and in some neurodevelopmental disorders.

A variety of microRNA expression profiling and genome-wide association studies have associated microRNA expression signatures and common polymorphisms at microRNA loci with a number of diseases, particularly cancers.

Observed associations include rs7372209 in miR-26a with decreased risk of bladder cancer in females[187] and rs11614913 in miR-196a with increased risk of breast cancer[188]. Levels of the microRNA processing machinery has also been associated with cancer. Dicer and Drosha were found to be decreased in ovarian cancer, with higher expression associated with increased median survival[189]. Expression profiles of microRNA genes have also been used as a biomarker to classify cancers with some prognostic utility[190].

Neurodevelopmental disorders have also been associated with microRNAs. Genome-wide association studies have identified miR-137 and SNPs near miR-137 target sites in schizophrenia[191]. Variants in miR-182 have also been suggested to affect major depression and insomnia through regulation of the circadian clock modulator (CLOCK gene) by miR-182[192].

A heterozygous deletion of the microRNA-17-92 cluster was found in a patient with

Feingold syndrome – a rare autosomal dominant disorder characterised by microcephaly, short stature and digital abnormalities – with targeted deletion of this microRNA cluster in mice phenocopying several features of the disease[193]. This microRNA cluster is hypothesised to be a downstream effector of MYCN, with loss of function mutations in MYCN responsible for over two thirds of Feingold syndrome cases[193].

1.7.4 Disease associated variants affecting microRNA expression

There are several examples of human disease associated or implicated variants that have been shown to impact microRNA expression:

A germ-line pri-microRNA mutant 7bp downstream of the miR-16 pre-microRNA was seen in two patients with chronic lymphocytic leukaemia (CLL), low levels of this microRNA having been identified as associated with CLL. Pri-microRNA constructs with this mutation expressed the microRNA at significantly lower levels than wild type[194] though the mechanistic basis for that difference in expression has not been demonstrated.

A common SNP rs2910164 within the miR-146a pre-microRNA associated with colorectal cancer risk (OR:1.34 95% CI 1.15-1.67) and survival (Hazard Ratio 2.12)[195] and breast cancer risk (OR:1.77, 95% CI 1.40-2.24)[196] was found through microRNA processing assays to reduce the level of pre-microRNA produced from the primary transcript during nuclear processing leading to a reduced level of mature microRNA[197].

A rare novel variant within the terminal loop of the miR-30c-1 pre-microRNA in a non-BRCA1/BRCA2 mutant breast cancer patient was predicted to alter the secondary structure of the precursor. This variant was found to increase the levels of the mature microRNA, a predicted regulator of BRCA1[198]. Investigation of the mechanism of this effect by the Caceres' lab (Edinburgh) found that the variant affected processing by the microprocessor complex. The effect of this variant on the RNA structure of the primary microRNA was then assayed using two complementary methods: Selective 2'-hydroxyl acylation analysed by primer extension (SHAPE), able to distinguish single stranded base-pairs which are highly reactive, from double stranded base-pairs which are non-reactive. And hydroxyl radical cleavage footprinting assaying the solvent accessibility of each nucleotide from the effect of hydroxyl radicals which break the RNA backbone in accessible regions. Combining these analyses they found that the variant present in the terminal loop caused a conformational change in the pri-microRNA, modifying the base pairing in both their terminal loop and in the basal stem. Assaying the effect of the variant on protein binding to the pri-microRNA through RNase-assisted chromatography followed by mass spectroscopy they found the variant led to binding of SRSF3, a factor previously described in microRNA biogenesis as binding to the CNNC motif downstream of the hairpin. Further validation trough SRSF3 overexpression and mutations of the CNNC motif confirmed that this interaction of SRSF3 at the CNNC motif was affected by the RNA structure changes caused by this variant in the terminal loop[199].

microRNA mutants have also been seen to cause a Mendelian developmental disorder. Separate mutations at adjacent base-pairs in the seed region of miR-96 were identified in two families as segregating with autosomal dominant deafness[200]. This microRNA is expressed in the inner ear, necessary for the maturation of the hair cells which transduce sounds into electrical stimuli[201]. Both of these mutants present in the seed region and therefore affecting microRNA targeting were also seen to affect microRNA biogenesis, with both mutations leading to an $\sim$80% decrease in expression of the mature microRNA[200].

1.8 Aims

Through this introduction I have highlighted the importance of microRNAs in the regulation of gene expression (Section 1.5), detailed the molecular steps in the biogenesis of microRNA (Section 1.3.1) and demonstrated that genetic variation affecting microRNA regulation can lead to human disease (Section 1.7.3).

Despite these insights, it is apparent that our understanding of microRNA mediated regulation is incomplete. Argonaute proteins are abundant in the nucleus and have clearly defined roles in the nucleus of plants (Section 1.5.3) but their roles and targets in human nuclei are not well defined.

The determinants of microRNA targeting, though clearly based in part on complimentary base-pairing are not fully understood so cannot be well predicted (Section 1.4.2). Similarly the consequences of sequence changes in the mature microRNA or its precursor cannot currently be predicted either in terms of processing efficiency or alterations in target specification (Section 1.7.2).

The work presented in this thesis sets out to use integrated genome-wide measures of microRNA-target interactions and expression data to systematically explore the nature of the microRNA mediated targeting of Argonaute, their sequence and context determinants and to understand how sequence changes to microRNA and their precursors can disrupt either the targeting or biogenesis of microRNAs. The specific aims that I will seek to address are:

- To use CLASH data to examine the genome-wide distribution of Argonaute targeting, identifying enrichment of microRNA targets amongst the transcriptome. Also examining the evidence for Argonaute targeting within the nucleus affecting transcription or splicing, and modelling microRNA targets genome-wide to examine the evidence for competitive endogenous RNAs.
- To examine the factors which contribute to microRNA processing using polymorphisms in human population. In both normal samples where these microRNA eQTLs can be identified in available genome sequencing and small RNA expression data. And in disease samples through exome and genome screening to determine if polymorphisms at microRNA loci contribute to disease phenotypes.

Chapter 2

Methods

This chapter describes datasets and software used in the following chapters, each subsequent chapter also describes in more detail the specific application of these methods.

2.1 Datasets

2.1.1 CLASH data

Nine independent cross-linking ligation and Sequencing of Hybrids (CLASH) datasets from Helwak et al[101] performed with slightly differing protocols are used to examine microRNA targeting in chapters 4 and 5.

In these CLASH experiments (schematic in Fig 2.1) HEK293 cells stably expressing N-terminally His tagged AGO1 proteins were UV irradiated to crosslink proteins with interacting RNAs. Tagged AGO1 proteins were purified and interacting RNAs were partially hydrolysed, ligated, reverse transcribed and sequenced. A subset of these sequenced reads (2-10% in the differing protocol variations) are hybrid reads representing intermolecular RNA–RNA interactions.

2.1.2 RNAseq datasets

Whole-cell and cell fractionated nuclear and cytosolic data were generated for this project and other Taylor lab projects, processing and analysis of this data are described in chapter 4, as well as its use in normalising for transcript abundance.

Whole-cell and cell fractionated nuclear and cytosolic data in five different cell lines from the ENCODE consortium[203] are compared with the generated HEK whole cell and cell fractionated data.

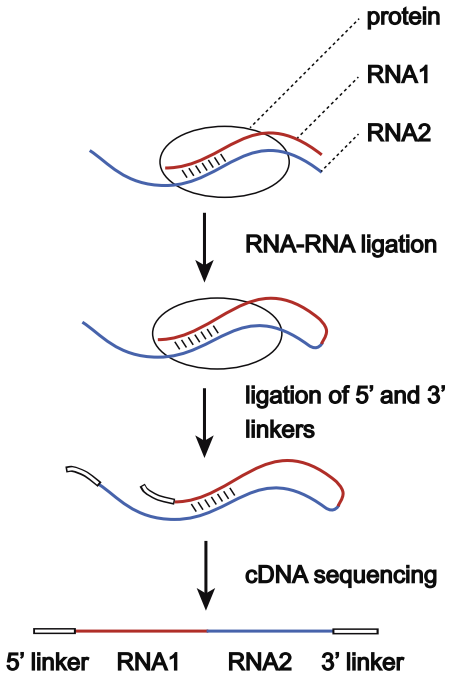


Figure 2.1: Schematic of the CLASH method showing purification, ligation and sequencing of protein bound RNA-RNA molecules. Applied to Argonaute protein a microRNA is shown in red, target RNA in blue.
Adapted from Travis et al.[202] licensed under CC BY 3.0.

2.1.3 eQTL datasets

Genome-wide DNA sequences from the 1000genomes project[204] and shortRNA sequencing performed in cell lines from the 1000genomes project by the GEUVADIS consortium[186] are used to identify microRNA eQTLs in chapter 6.

2.1.4 Patient exome data

A variant database from exome sequencing and variant calling of 1295 patients or parental controls from five patient cohorts within the IGMM are analysed in chapter 7 examining variants at microRNA loci.

2.1.5 Other resources

- All alignments and annotations are made to the GRCh37 genome assembly.
- GENCODE[205] v19 annotations are used throughout
- RepeatMasker[206] v3.3.0 (2012/01/24 hg19 annotations) are used in chapter 4
- miRBase[11] v20/v21 microRNA annotations are used throughout.

- FANTOM5[207] transcription start site annotations are used in chapter 4.
- Predicted circular RNAs from Memczak et al[160] are used in chapter 4.
- SNPs from dbSNP[208] v139 and ExAC[209] release 0.3 variants are used for comparison throughout.
- DECODE[210] derived allele frequencies are used to test for selection in chapter 6.
- TargetScan[96] v7 predicted microRNA targets are used in chapter 7.

2.2 Software

hyb[202], a pipeline developed with collaborators for the identification of hybrid sequence reads and applied to CLASH data is described in chapter 3 and is available at https://github.com/gkudla/hyb.

I have also developed hyb_gen, an extended hybrid read identification pipeline for alignment and analysis genome-wide, described in chapter 3.

2.2.1 Programs

The following programs have been used in this thesis:

- Bowtie2[211] v2.1.0: genome alignment
- Tophat2[212] v2.0.10: Splice aware aligner
- blastn[213] v2.2.26: aligner
- UNAfold[214] v3.8: RNA folding energy calculation
- STAR[215] v2.4.2a: Splice aware aligner
- Kallisto[216] v0.42.3: Transcript quantification
- DEseq2[217]: Quantification normalisation
- sleuth[216]: Quantification normalisation
- BEDTools[218] v2.21.0-26: Operations on genomic intervals
- BCFtools[219] v1.2-5: Operations on variant data
- H3M2[220]: Copy number analysis in exome data
- Forgi v0.2: An RNA manipulation python package (http://www.tbi.univie.ac.at/∼pkerp/forgi/)
- python v2.7.1: Numerous custom scripts were written in python

- R[221] v3.2.2: Datasets were analysed and graphed using R and the packages below

2.2.2 R packages

- ggplot2[222] v2.1.0: Graphs
- data.table[223] v1.9.7: Data manipulation
- MatrixeQTL[224] v2.1.1: eQTL testing
- VennDiagram v1.6.16: Venn diagrams
- gplots v2.17.0: for heatmap.2 function
- party[225] v1.0-25: cforest function for random forest implementation.

Chapter 3

Hybrid read mapping pipeline

3.1 Introduction

Inter-molecular RNA–RNA interactions are key to many cellular processes including; pre-mRNA splicing where spliceosomal RNAs bind to splice site motifs, ribosome synthesis where snoRNAs guide the modification of other RNA molecules, translation where tRNA molecules base pair with complementary mRNA codons, and the activity of microRNAs, which bind to other RNA molecules in the cell through complementary base pairing (section 1.4), leading to the suppression of protein translation from a target mRNA or facilitating the degradation of the target (section 1.5).

A number of methods have been developed to assay microRNA targets in cells, generally extensions of the CLIP protocol where UV exposure is used to crosslink proteins with RNA, proteins of interest can selected for, and pools of RNA can be further processed and sequenced in high-throughput (section 1.4.3).

The CLASH protocol[100] is one extension of CLIP where after selection for the protein of interest an RNA ligation reaction is performed. Applying this to Argonaute attempts to ligate the microRNAs bound by Argonaute to their target RNA molecule. The pool of RNA molecules sequenced will then contain some fraction of hybrids, one part microRNA and one part target. The identification of these hybrid RNA sequence reads and resolution of the component interaction partners is then a computational problem to be addressed.

One solution to this hybrid read mapping problem was applied by Helwak et al. to the first CLASH experiments identifying 18,500 microRNA-mRNA interactions[101]. In a collaboration with the authors of that original solution, I have tested, revised, and produced a revised pipeline as a standalone package named hyb[202]. Specifically I have implemented: alignments using Bowtie2, a substantial improvement in speed over previous versions using blastn. Tested the hyb package against Tophat2-fusion[226] a popular method designed for calling gene fusions with hybrid reads. Tested and debugged the standalone package for bugs

and improvements to usability.

A limitation of the hyb based mapping approach and other previously developed microRNA:target mappers was that they assumed the targets were annotated transcripts. As much of the genome is known to be transcribed[227] and Argonaute mediates RNA interactions in the nucleus as well as the cytoplasm (section 1.5) there is a strong rationale for the development of a genome wide CLASH mapping algorithm that can explore the possibility of non-canonical Argonaute mediated interactions and also for the unbiased investigation of all microRNA interactions in less well studied transcriptomes (cell types, cancer genomes or species) where there is a genome sequence or appropriate reference available.

Building on hyb I have developed a separate pipeline using genome-wide mapping to investigate the diversity and distribution of microRNA targets, this creates some additional challenges including accounting for multi-copy and highly repetitive sequences and the separation of transcripts into exons. This pipeline has been used in chapters 4 and 5.

3.2 hyb - development and testing

The original version of hyb was developed by Grzegorz Kudla for the analysis of CLASH data, the outline of the hyb pipeline is shown in Fig 3.1. It used the local alignment of reads to a transcriptome database with blastn[213], these alignments were then processed to identify hybrid reads. Where non-end-to-end alignments are found hybrids were identified through an iterative process starting with the top scoring match comparing each additional alignment in score order for whether they map the remaining portion of the read allowing a gap or overlap of 4bp. Where there were multiple hybrids possible the selection criteria were: 1) the sum of the alignment scores for both fragments of the chimera. 2) The transcript classes of the chimera favouring microRNA-mRNA hybrids. 3) The transcripts with the highest total number of reads.

Features which were identified to be improved on from this original version for the publication of a standalone package were the run-time, changing the aligner to make use of read quality, and incorporating the option for a genomic alignment.

3.2.1 Bowtie2 alignment

Alignments in hyb were initially performed with blastn[213], which while allowing sensitive alignment of read fragments was time consuming for large datasets and was not able to take advantage of the FASTQ read quality information.

Bowtie2[211] a popular alignment tool for next generation sequencing data has an option for local alignment, allowing fragments of reads to be aligned rather than whole reads being aligned end-to-end as in the original Bowtie.

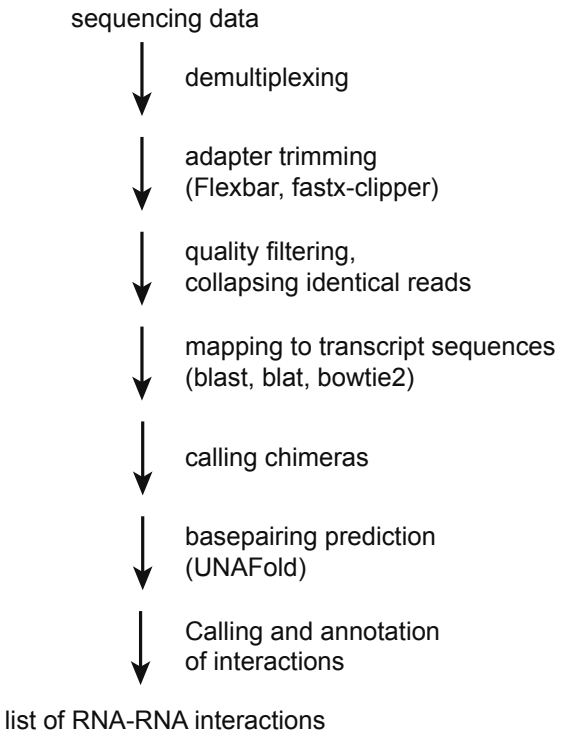


Figure 3.1: Steps performed by the hyb analysis pipeline
Adapted from Travis et al.[202] licensed under CC BY 3.0.

Using a list of reads which had been identified as hybrid reads after blastn alignment and processing using an e-value threshold of 0.1 with all other parameters default, I determined via trial and error a set of Bowtie2 parameters to approximately reproduce these results using the built in `--very-sensitive` parameter set to control the number of seed extension attempts with a smaller seed length (16bp) reporting alignments with a minimum score of 18.

Applied to the 'E4' CLASH dataset, $\sim$34 million reads from one variation of the CLASH protocol performed by Helwak et al[101], mapped to a custom transcriptome database using Bowtie2 v2.1.0[211] and blastn v2.2.26[213] identified similar numbers of hybrids: 14019 in common, 3313 only in blastn and 2711 only with Bowtie2. In the absence of a complete set of known interactions it would be useful to have some orthogonal measure of sensitivity and specificity for the prediction of these interactions. One such measure that has been used previously is that of folding energy[228]. More negative folding energies between microRNA and target indicate greater nucleotide complementarity and stability of a hydrogen-bonded duplex of the RNA molecules, thought to be an important component of the interaction between microRNA and target. Hybrids identified following alignments with both programs both had the same folding energy distribution with median value $\sim$-18kcal/mol. However the run time with Bowtie2 was $\sim$1 hour, instead of $\sim$11 hours with blastn.

3.2.2 Tophat2 fusion-search comparison

As a comparison to an existing method hyb was compared with the Tophat2 v2.0.10 fusion-search option[212], developed for the identification of fusion transcripts in cancer. This is a solution to a similar problem based initially on finding reads with segments aligning to different transcripts or to different genomic locations. Tophat2 fusion-search uses a multi stage process where reads are first mapped end-to-end, unmapped reads are then cut into fixed width segments and each aligned end-to-end, those segments from the same read which aligned on separate chromosomes or greater than 100kb apart are then used to form 'fusion contigs' which the initially unmapped reads are aligned to. Read coverage filters can then be applied to define fusion events.

The parameters for Tophat2 fusion-search had to be modified to allow the identification of short microRNA segments, and to not impose a read coverage threshold. The results of both pipelines were compared for; the number of microRNA hybrids identified, and the folding energy measuring how well the two RNA fragments base pair calculated with hybrid-min from the UNAfold v3.8[214].

With optimised parameters for both Tophat2 v2.0.10 fusion-search and the updated hyb, Tophat2 fusion-search identified 8231 microRNA:mRNA hybrids and hyb identified 13483. This raises the question as to whether hyb is over-predicting or Tophat2 fusion-search is under-predicting interactions.

As neither hyb nor Tophat2 fusion-search makes use of folding energy in their identification of hybrids, this can be used to estimate the relative enrichment of genuine Argonaut mediated microRNA:target interactions[94]. Tophat2 fusion-search microRNA-mRNA hybrids had a mean folding energy of -10.3kcal/mol, whereas the hyb identified microRNA-mRNA hybrids had a mean folding energy of -18.2kcal/mol, the distribution of these folding energies is shown in Fig 3.2 compared to a shuffled dataset where microRNA-mRNA pairs are randomised. This suggests that Tophat2 fusion-search misidentified many hybrids, possibly due to the fixed with segments of reads which are aligned to produce the 'fusion contigs' rather than the local alignments used by hyb.

3.2.3 Usability and testing

hyb is implemented as a makefile calling additional programs and supplied scripts to generate the desired outputs. Combinations of varying types of input file and processing options were tested to ensure expected results, useful help and error messages, and specification of prerequisites.

I have published hyb as a standalone pipeline with my collaborators [202], it is publicly available at https://github.com/gkudla/hyb distributed under the GNU GPL (General Public License).

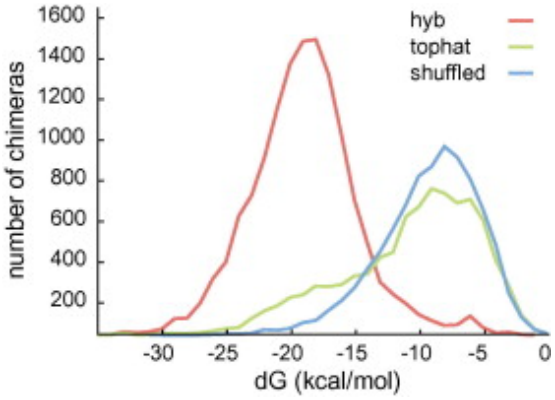


Figure 3.2: Comparison of hyb and Tophat2-fusion in the identification of hybrid sequence reads. Distribution of folding energies of microRNA-mRNA chimeras recovered with hyb, Tophat2 fusion-search, and in randomly re-associated microRNA-mRNA pairs from the Tophat2 fusion-search analysis. More negative dG values represent stronger base-pairing.

3.3 Genome-Wide mapping pipeline

I have created an alternative hybrid mapping pipeline using alignment to a reference genome, this has the advantage of ensuring that reads are aligned to their best match regardless of transcript type, allowing analysis of their genome-wide distribution. Some aspects of this such as using Bowtie2[211] for fast local alignment have since been incorporated into the hyb package.

3.3.1 Overview of the pipeline

After trimming sequencing adaptors, reads are aligned to the genome using Bowtie2 local alignment as described previously (section 3.2.1), the resulting sam file is then processed to identify hybrids based on a set of criteria:

- Reads must not align to more than **m** locations in the genome above an alignment threshold **n**. Highly repetitive sequences which cannot be uniquely mapped are of limited use in downstream analysis and will require excessive memory and processor use to calculate all such mapping locations and examine them for hybrid pairs. Limiting this parameter will substantially improve performance.
- If the best scoring alignment does not cover the whole read, pairs of alignments are considered to identify if any two have a higher alignment score than the best single alignment, these are the candidate hybrid reads.
- To be considered a hybrid read the pair of alignments must cover more than 80% of the read length. This allows for some unmapped bases due to insufficient linker trimming, but ensures source of the read is a single hybrid of two RNA molecules.

- The pair of alignments must also not have a gap or overlap $>\mathbf{p}$ bases. A gap/overlap will allow for ambiguous edges, where the alignments may by chance extend into the partner.

If reads can be assigned to more than one hybrid with the same alignment scores they are assigned to the hybrid with the highest number of non-hybrid reads calculated by summing the non-hybrid reads which overlap each potential hybrid. Such that hybrids with ambiguous alignments are assigned to the alignment with the most supporting evidence (a rich get richer strategy).

The final list of unambiguous hybrids (one read: one hybrid) are then be collapsed using a custom python script such that 10 reads hybrid for locusA-locusB will appear once with a count of 10. These lists are then be intersected with regions of interest using tools such as BedTools [218] and the folding energy between the RNAs of each hybrid can be calculated using UNAfold [214] or the Vienna package[229].

3.4 Parameter sweep

A parameter sweep was performed for this pipeline, running the pipeline multiple times on the same dataset with a range of parameters to identify optimal values based on a scoring criteria. The E4 dataset of Helwak et al.[101] was used, varying the alignment score threshold (16,17,18,19,20) and the hybrid calling parameters: maximum number of alignments per read (5,10,20), and allowed gap/overlap between the alignments (0,1,2,3,4,5,6). Performance was measured by:

- The number of microRNA containing hybrids identified, to be maximised as these are the primary data for downstream analysis.
- The mean folding energy of those microRNA hybrids as calculated by UNAfold v3.8[214], this score is to be minimised, where lower folding energy corresponds to higher complementarity between microRNAs and their targets. As no true positive set of hybrids is known this orthogonal score not used in the generation of hybrids is used as a proxy for their sensitivity/specificity.
- The number of mRNA-mRNA hybrids identified which correspond to true splice variants, identified as reads which align to an mRNA transcriptome database but as hybrids when aligning to the genomic database. These are a different set of hybrid read expected to be present in the database when using a genomic alignment. This is to be maximised as in this case a set of true hybrids is known, although mRNA-mRNA hybrids may have different properties to microRNA-mRNA hybrids.

3.4.1 Results

Fig 3.3 shows the results of this parameter sweep. Optimum parameters were selected based on maximizing the number of microRNA-mRNA hybrids called and true exon-exon splice variants while minimizing the mean folding energy of the microRNA-mRNA hybrids. Relaxing thresholds lead to identifying more hybrids but with a lower mean folding energy. Parameters selected to identify hybrids for use in chapters 4 and 5 were alignment score 19, gap/overlap 4 and alignments per read 10. These parameters may be easily configured by the user for application of the pipeline in different contexts, e.g. with longer sequencing reads.

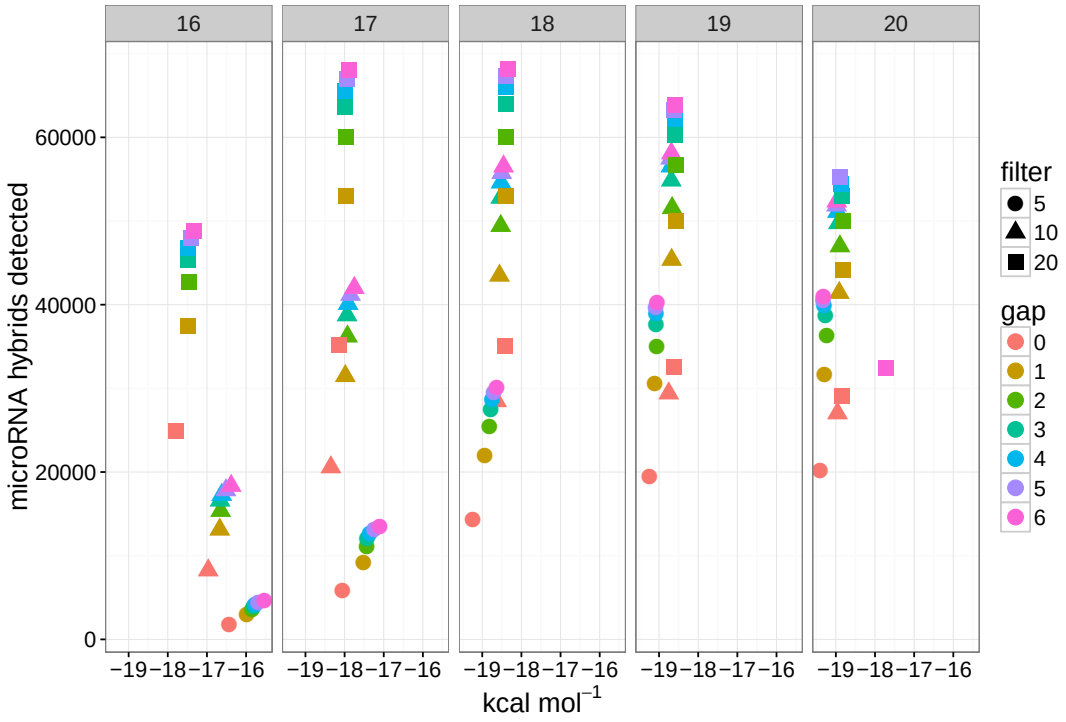
3.5 Chapter summary

I have published a pipeline - hyb - for the identification of hybrid sequencing reads[202]. Developed for the analysis of CLASH data it has been extended, tested against a popular fusion mapping program and made freely available online.

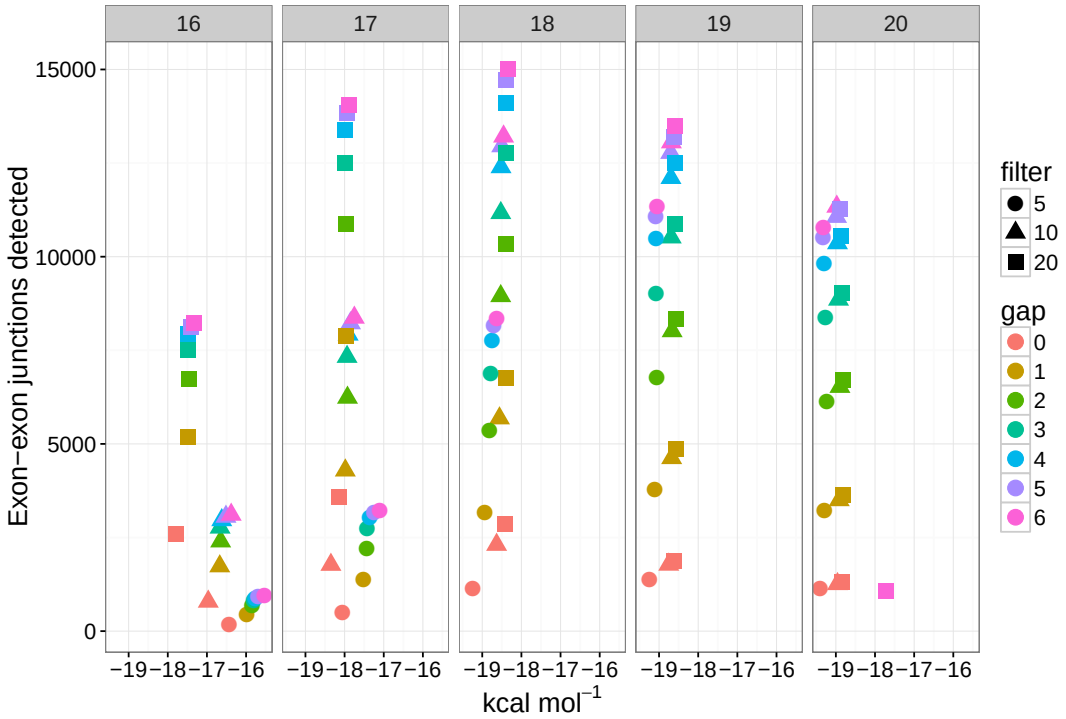
I have also created a pipeline for the analysis of CLASH data genome-wide. Some aspects of this pipeline such as the use of Bowtie2[211] local alignment allowing genome wide alignment within a reasonable time frame for large datasets have subsequently been incorporated into the hyb pipeline. This genome-wide hybrid mapping pipeline was optimised to identify microRNA-mRNA hybrids and exon-exon junctions in a CLASH dataset while maintaining a low minimum free energy in the folding of the microRNA-mRNA hybrids.

Applying this pipeline to AGO1 CLASH data allows the analysis of small RNA-mediated RNA–RNA interactions identifying both segments. These hybrids will be analysed in the following chapters; chapter 4 examining the genome-wide distribution of hybrids and their enrichments in classes of transcripts, and chapter 5 examining clusters of hybrid targets corresponding to sites of high occupancy including competitive endogenous RNAs.

This pipeline has also been applied in other projects at the IGMM, being used to identify genomic rearrangements at specific loci in whole-genome sequencing data[230].



(a)



(b)

Figure 3.3: Results of parameter sweep. (x-axis both) mean folding energy of microRNA-mRNA hybrids. (y-axis a) number of microRNA-mRNA hybrids. (y-axis b) number of recovered exon-exon junctions. (panel number) alignment score threshold. (colour) gap/overlap threshold. (shape) alignments per read threshold. Ideal results would be towards the top-left of the graph i.e. many strong interactions.

Chapter 4

Genome-wide microRNA targets

4.1 Introduction

In the post-transcriptional regulation of protein coding genes by microRNAs, microRNAs as part of the RISC complex bind to the 3'UTRs of protein coding genes leading to their translational repression or cleavage (section 1.5). However a number of studies have identified microRNA targets within exons and lincRNAs hypothesised to act as competitive endogenous RNAs limiting the amount of microRNAs free to repress their gene targets[159, 160, 161] (section 1.6). Studies have also suggested microRNA targeting in the promoter or internal exonic regions with hypothesised roles in transcription[137, 231, 139, 132, 144, 147, 141] (section 1.5.3) or splicing[154, 153, 151, 150] (section 1.5.4).

In this chapter AGO1 mediated RNA–RNA interactions are identified in all 9 CLASH datasets from Helwak et al[101] using the genome-wide hybrid mapping pipeline discussed in chapter 3. These RNA–RNA interactions are analysed for the distribution and specificity of microRNA targeting.

One important factor in analysing the distribution and specificity of microRNA targets is transcript expression level in assessing target specificity and location, as hypothesised effects on transcription or splicing must occur within the nucleus.

In order to address this I have used cell-fractionated cell-matched RNA-seq to provide estimates of transcript abundance and sub-cellular localisation. This allows me to test the widely held hypothesis that microRNAs primarily target the 3'UTR of protein coding transcripts and that these interactions occur primarily in the cytoplasm.

4.2 Methods

4.2.1 Datasets

Nine independent cross-linking ligation and Sequencing of Hybrids (CLASH - protocol described) datasets from Helwak et al[101] performed with slightly differing protocols were used in this analysis (section 2.1.1).

RNAseq in whole cell and cell fractionated (nuclear/cytosolic) HEK293 cells was commissioned for use in this project and other Taylor lab projects. HEK293 cells from IGMM stocks were maintained in DMEM supplemented with 10% fetal calf serum and tested negative for mycoplasma contamination. Cell fractionation and sequencing was performed for two cell culture replicates with each fraction spread over multiple lanes allowing for the detection and correction of various batch effects.

4.2.2 Mapping with genomic hyb

Mapping single and hybrid reads was performed with the genomic hyb pipeline discussed in chapter 3, using the default parameters selected by the parameter sweep for optimal performance in whole mammalian genome alignment.

4.2.3 Annotation

Hybrid read segments were annotated by intersection with a bed file comprised of miRBase v21[11] mature microRNAs, repeatmasker[206], Gencode v19[205], and manual annotations of rDNA loci. An annotation hierarchy was used where multiple annotations overlapped a hybrid read segment in the order of: rDNA, microRNAs, repeat loci, gene 3'UTR, gene 5'UTR, gene exon, gene intron.

Folding energy

The folding energy between the hybrid segments was calculated using hybrid-min from the UNAfold v3.8[214] package, as a measure of how well the RNA molecules base-pair together.

RNAseq quantification

In order to normalise to RNA expression level, RNA sequencing data to high depth from HEK293 cells was commissioned and analysed. Consisting of Whole cell and cell fractionated (nuclear / cytosolic) each with two biological replicates spread across 6 lanes of sequencing giving 36 fastq files per sample. Alignment was performed with STAR v2.4.2a[215]

with parameters used by the ENCODE project[203], which also output read counts for Gencode transcripts. Read counts were analysed with DEseq2[217] performing library size normalisation and tests of differential expression.

Transcript quantification was also performed with the alignment free program Kallisto v0.42.3[216], these results were compared to quantifications after STAR alignment to verify the alignment free approach in this and future projects. Transcript quantification with Kallisto v0.42.3[216] was also performed in the same way for five cell types from the ENCODE project[203] with similar whole cell and nuclear/cytosolic cell fractionated data.

Transcripts and genes are then quantified using TPM - transcripts per million, interpreted as if you were to sequence 1 million full length transcripts from a sample, the number of transcripts of one type seen is given by its TPM. This measure proportional to RPKM/FPKM within an experiment is more consistent between experiments[232]. Quantifications were made to an index comprised of the protein-coding and long non-coding RNA transcripts from Gencode v19[205]. Quantifications were normalised for library size using the `sleuth_prep` and `kallisto_table` functions from the sleuth R package[216].

These transcript quantifications are used in this project and other Taylor lab projects.

Statistical modelling

Linear models were used to examine the influence of transcript expression level on CLASH targeting by performing regression with second order polynomials using the R functions `lm` and `poly` from the stats package.

A random forest[233] algorithm based on conditional inference trees using the R function `cforest`[225] from the party[225] v1.0-25 package and a variable importance measure using the R function `varimpAUC`[234] also from the party[225] v1.0-25 package were used to build models for and estimate the importance of different predictors in predicting CLASH targeting.

Statistical tests

Enrichment analyses for microRNAs targeting specific locations were performed with Fisher's exact tests using the R function `fisher.text` from the stats package.

4.3 Results

4.3.1 RNAseq quantification

Table 4.1 shows the RNA-seq sample information for the commissioned cell fractionated HEK samples. Each sample replicate was run in triplicate split across six lanes of two sequencing

run	lane	replicate	cytosol	nucleus	whole.cell
110	1	1	3	3	3
110	1	2	3	3	3
110	2	1	3	3	3
110	2	2	3	3	3
257	5	1	3	3	3
257	5	2	3	3	3
257	6	1	3	3	3
257	6	2	3	3	3
257	7	1	3	3	3
257	7	2	3	3	3
257	8	1	3	3	3
257	8	2	3	3	3

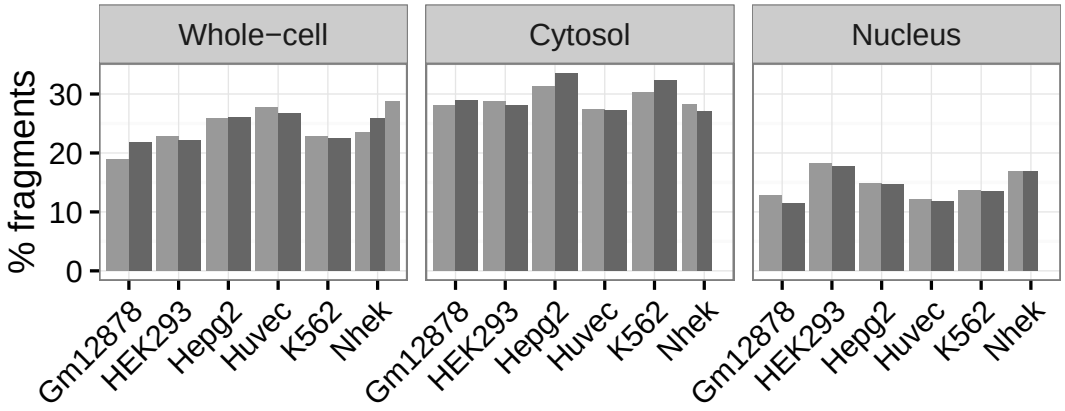
Table 4.1: HEK293 RNAseq experimental design: number of cytosolic, nuclear and whole cell samples in each sequencing run, lane, and culture replicate

runs. Each of the 18 samples run for each replicate had between 8.5 million and 19.2 million reads (median 13 million), giving sample a combined number of reads between 211 million and 270 million (median 259 million). This experimental setup allows for the detection and correction of various batch effects.

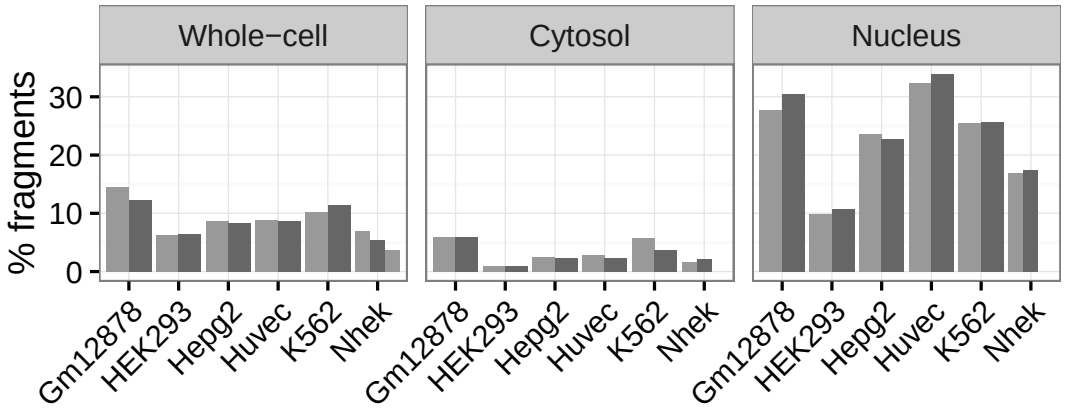
After alignment with STAR[215] v2.4.2a the HEK samples had median 91% reads uniquely mapping, with the ENCODE[203] samples having median 88% reads uniquely mapping. The proportion of reads aligning to constitutive exons and introns was calculated for both the HEK and five ENCODE cell type datasets, the results shown in Fig 4.1. These suggest that the HEK fractionation was successful with exonic fragments highest in the cytosol and lowest in the nucleus, and the reciprocal result for intronic fragments. Compared with the ENCODE cell types, the proportions of reads in the whole cell fraction seem broadly similar, in the cytosol fraction they are again broadly similar with the HEK data having very few intronic reads due to little nuclear contamination, the nuclear fraction seems somewhat different with more exonic and fewer intronic reads than the ENCODE cell types suggesting some cytoplasmic contamination. These differences between the HEK and ENCODE cell types will likely reflect that different cell fractionation protocols were used to generate them.

Recently alignment free methods for RNAseq quantification have become more popular due to their speed and proposed accuracy. I have processed the same HEK RNAseq datasets using the alignment free method Kallisto[216] v0.42.3. Fig 4.2 shows comparisons for estimated fragment counts from Kallisto and actual fragment counts from STAR[215] v2.4.2a, with spearman rank r^2 values greater than 0.95 in all three cellular fractions. Length normalised estimated fragment counts from Kallisto (TPM - transcripts per million) were used for transcript normalisation in this chapter.

Examining the overlap in expression levels in the cellular fractions in Fig 4.3, the numbers of genes expressed with TPM >1 and their overlaps are broadly similar between the cell types,

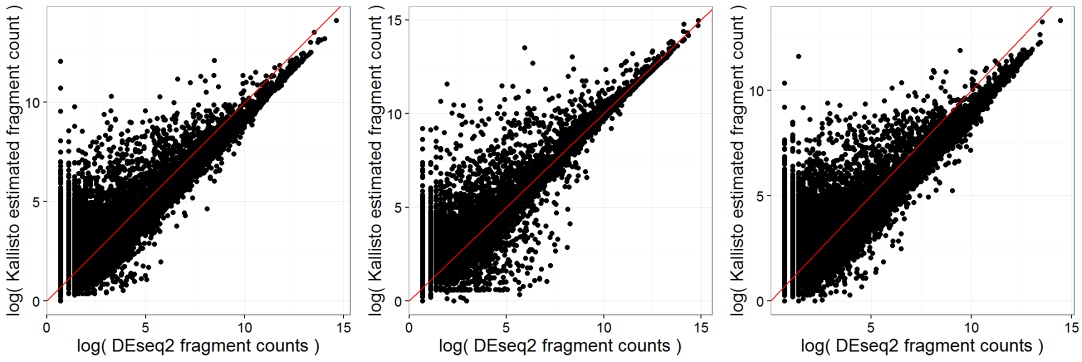


(a) Exonic reads



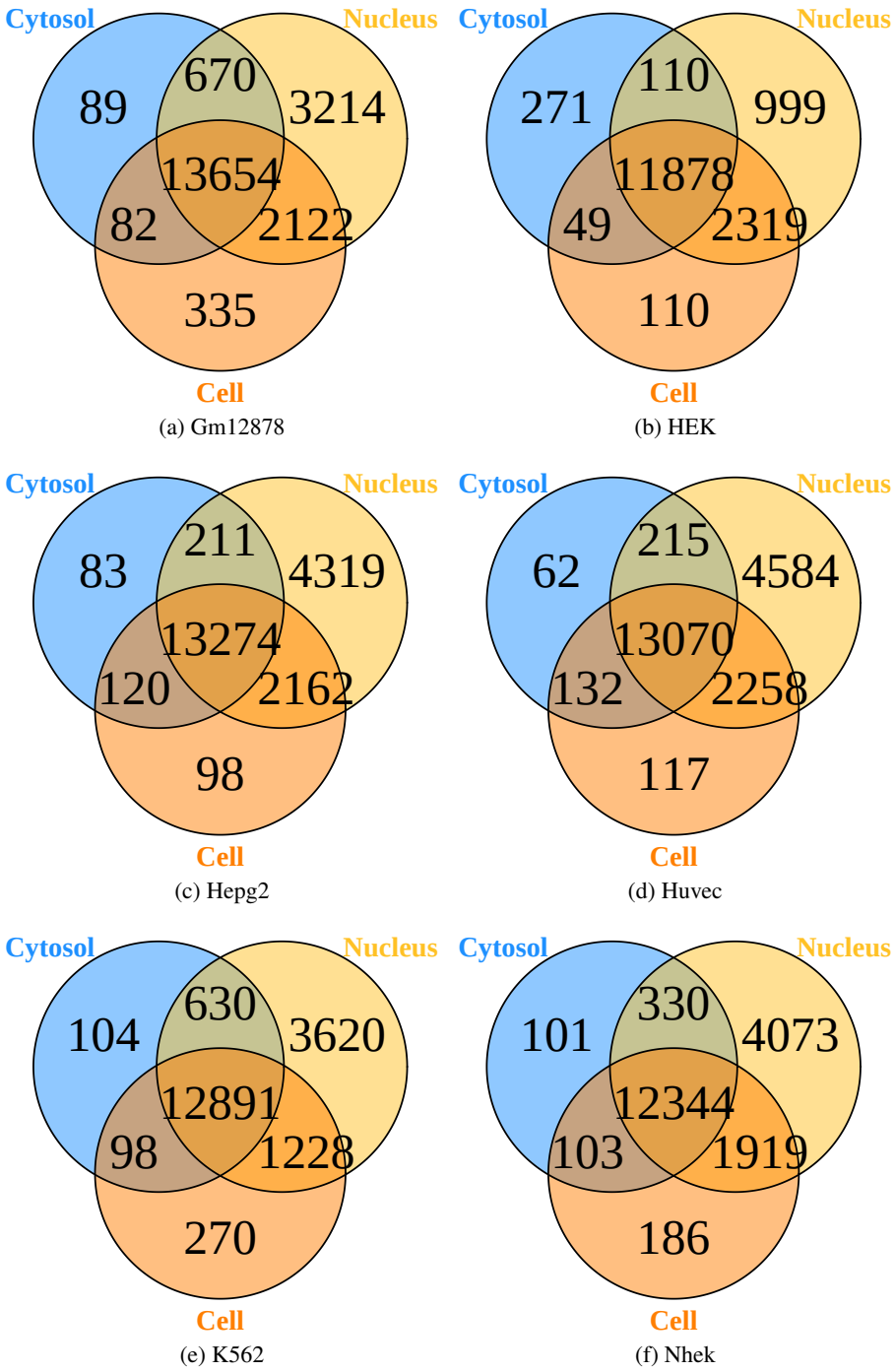
(b) Intronic reads

Figure 4.1: RNAseq fragments aligning to (a) exonic or (b) intronic regions. For different cell types, 5 from ENCODE and HEK293. Replicates are shown with differing shades of grey. Values indicate the success of the cell fractionation protocols.



(a) Whole cell r^2=0.952 (b) Cytosol r^2=0.964 (c) Nucleus r^2=0.960

Figure 4.2: Estimated fragment counts from the alignment free method Kallisto[216] v0.42.3 and STAR[215] v2.4.2a alignment. Spearman rank r^2 values shown.



(a) Gm12878 (b) HEK (c) Hepg2 (d) Huvec (e) K562 (f) Nhek

Figure 4.3: Venn diagrams displaying the numbers of genes expressed >1 TPM in the whole cell, cytosol, and nuclear fractions for the six different cell types. Five from ENCODE and the generated HEK data.

with fewer genes in the HEK cells fewer nucleus specific genes and more cytosol specific genes. As TPM is based on the relative proportion of transcripts in the sample this difference in the HEK cells may be due to additional non-nuclear transcripts contaminating the HEK sample, and fewer non-cytosolic transcripts contaminating the cytosolic fraction.

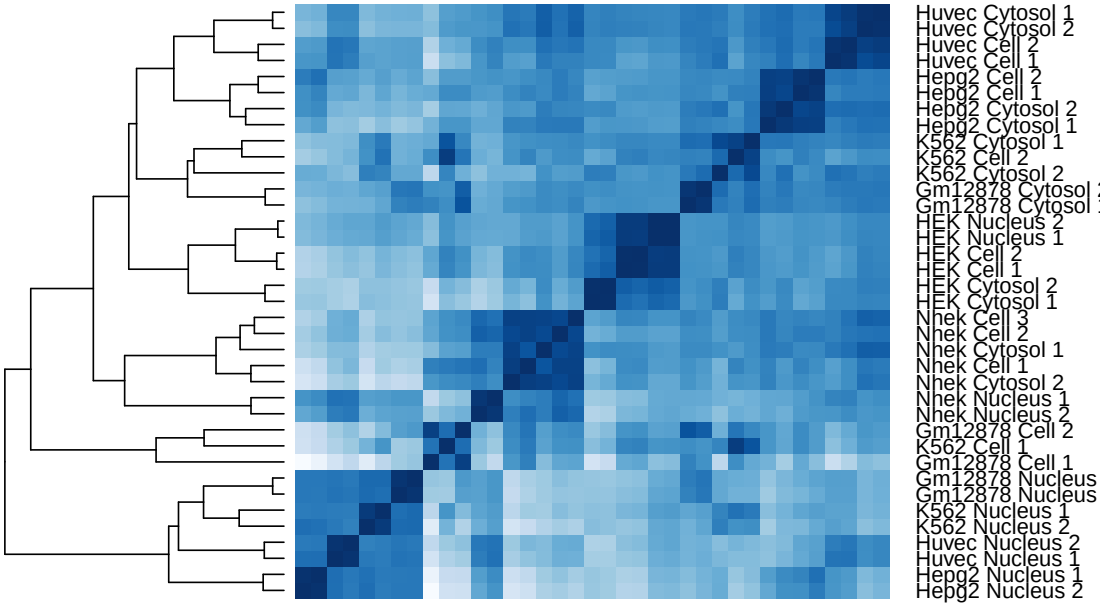


Figure 4.4: Heatmap displaying clustering of Spearman rank correlation of genes with TPM>1. Replicates cluster together, after this some cluster by subcellular fraction, others by cell of origin.

Clustering these expression datasets by Spearman rank correlation is shown in Fig 4.4. Replicate datasets cluster together as would be expected, beyond that some samples are clustering by fraction with the nuclear fractions of Gm12878, K562, Huvec and Hepg2 clustering together. While HEK and Nhek samples are clustering first by cell type, then by cellular fraction. The HEK and Nhek cells were also those with the lowest fractions of intronic fragments shown in Fig 4.1. Differences in the HEK and Nhek cells here could be due to contamination of non-nuclear transcripts during cellular fractionation or genuine differences between the distribution of transcripts in these cell types.

4.3.2 microRNA targeting - transcript diversity

40,600 microRNA targets were identified amongst all 9 CLASH datasets from Helwak et al. Table 4.2 shows the count of how many hybrids were identified in multiple experiments, with 78% being observed in just a single experiment.

The count of microRNA targets annotated to different classes of transcript is shown in Fig 4.5, with the highest categories being protein coding 3'UTRs followed by protein coding exons. Many hybrids are also seen between microRNAs and repetitive elements including rRNAs, LINEs, SINEs and tRNAs. Another large category of hybrids are those where the target sequence is not present in the annotation dataset labelled 'unknown'.

After normalising for the total genomic sequence with each annotation, the number of hybrids per megabase is shown in Fig 4.6. Here the category with the highest number of microRNA targets are microRNAs themselves present in the dataset as microRNA duplexes. Of these 1032

Number of experiments	hybrids
1	31738
2	6483
3	1054
4	563
5	336
6	210
7	119
8	67
9	30

Table 4.2: Number of CLASH hybrids identified in 1 or more CLASH experiments of Helwak et al, where the protocol was varied slightly. Most hybrids are seen only in one experiment

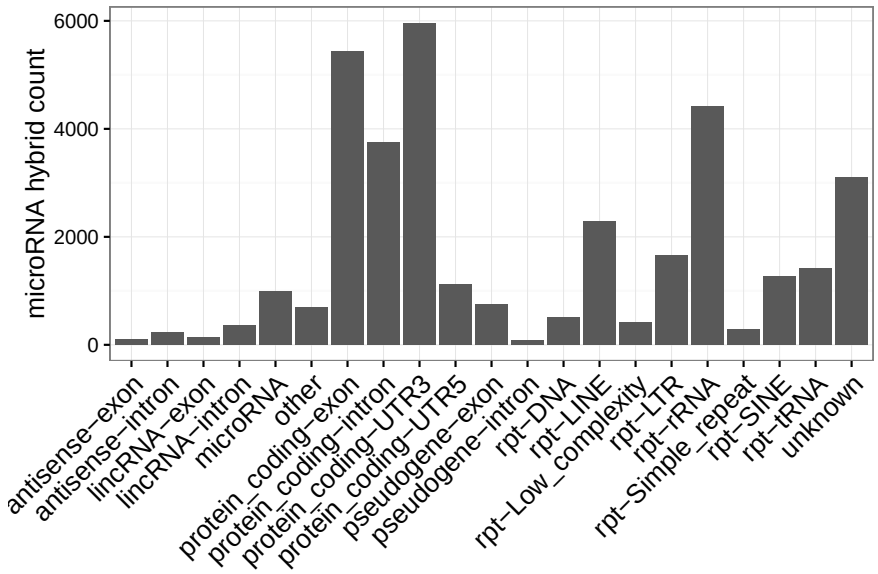


Figure 4.5: The count of microRNA hybrids annotated to different transcript classes. (X axis) Transcript annotation. (Y axis) Count of microRNA hybrids with transcripts of that class. The majority of transcripts targeted by microRNAs are protein coding genes along the full length of transcripts and repetitive sequences.

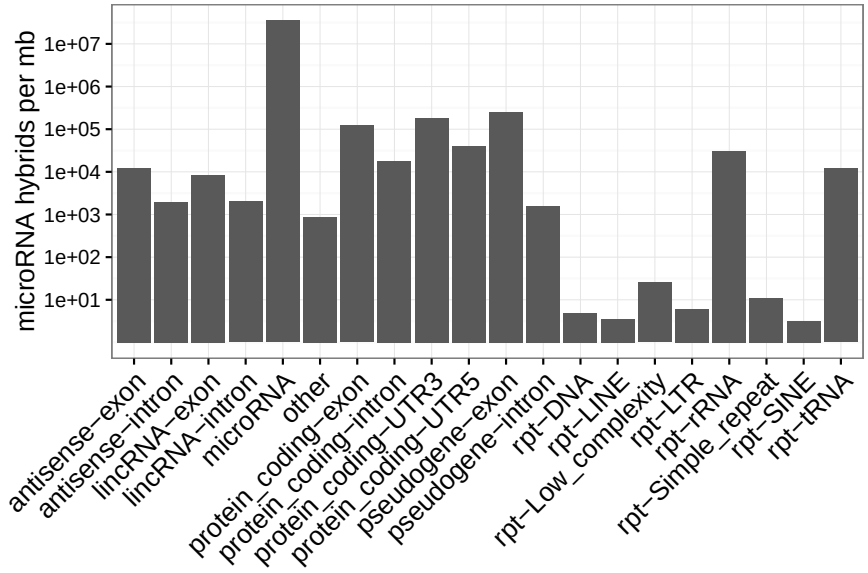


Figure 4.6: The count of microRNA hybrids annotated to different transcript classes normalised to the genomic length of the annotation. (X axis) Transcript annotation. (Y axis) Count of microRNA hybrids in transcripts of that class per megabase of the genome with that annotation. Normalised in this way protein coding genes, lincRNAs, pseudogenes, rRNAs, tRNAs and antisense transcripts display similar levels of microRNA binding, while microRNA duplexes are enriched.

microRNA duplexes 240 are pairs of mature microRNAs from the same precursor hairpin, and 792 are duplexes of mature microRNAs from different precursor hairpins. As the total amount of microRNA sequence in the genome is relatively small these microRNA-microRNA hybrids appear more enriched in this plot after normalising for genomic sequence. This binding of duplex microRNAs may represent an additional method that is used to control the level of free microRNAs available to repress their gene targets.

The number of microRNA hybrids present at protein coding loci after this normalisation seems more similar to other categories with rRNA, tRNA, lincRNA and antisense transcripts having similar numbers of microRNA hybrids per megabase within two orders of magnitude 1,000 - 100,000 hybrids per megabase.

The remaining repetitive loci show low enrichment of microRNA hybrids when normalised for genomic content, with LINEs, SINEs, low complexity and simple tandem repeats having fewer than 10 hybrids per megabase of genomic sequence. This may be due to a large fraction of the genome that these RNAs occupy, and their highly repetitive nature means they will have reduced mapability so mapping reads would likely be discarded by the pipeline. However the vast majority of these regions are likely to be non-transcribed, so this normalisation is likely conservative considering level of transcripts produced by these regions. In CLIP like experiments, alignments to these RNA species present at many copies in the genome are often not considered, but these result suggest they may represent a substantial proportion of microRNA targets.

4.3.3 microRNA target specificity

Normalising for transcript abundance using RNAseq data is possible for long transcripts captured by the RNA-seq protocol - protein coding genes, lincRNAs, and antisense transcripts. Fig 4.7 shows scatter plots comparing each transcript for the number of CLASH reads annotated to it and its expression level in TPM from whole-cell RNAseq, with quadratic lines of best fit and their r^2 values. Amongst the intronic categories a little relationship is seen between transcript abundance and the number of CLASH reads with r^2 values around 0. Antisense transcripts also had little relationship in either the intronic or exonic segments. Amongst the other categories; protein-coding, lincRNA, protein-coding 3'UTRs and 5'UTRs a positive relationship is seen with r^2 values ranging between 0.29 and 0.45. This positive relationship can be seen particularly for transcripts with expression values >1 TPM. In HeLa cells it has been estimated that a TPM of 1-10 corresponds to 1 copy per cell[235]. As intronic segments will not be retained in the mature transcript their expression is likely not well captured by this RNA-seq, however using expression data from the HEK nuclear fraction did not alter the relationship seen here for intronic targets.

Similarly seen in Fig 4.8 the folding energies corresponding to microRNA:target base pairing were stronger for exonic, 3'UTR or 5'UTR targets compared to intronic targets in both protein coding and lincRNA transcripts.

Another metric for microRNA target reliability is the presence of non-hybrid reads at the target locus as supporting evidence of a reliable interaction. Fig 4.9 shows the fraction of targets that also have non-hybrid reads at the targets locus for each annotation category. The intronic categories have the lowest fraction of supporting non-hybrid reads with $\sim$20-30% of targets, while protein coding 3'UTRs, 5'UTRs and exons have the highest fraction of supporting non-hybrid reds with $>$75% of targets. The remaining antisense exons, lincRNA exons, and other categories have $\sim$50% of targets with supporting non-hybrid reads.

Of the 7431 Gencode v19[205] transcripts which contain microRNA targets identified as hybrid CLASH reads only 251 transcripts contain multiple targets from the same microRNA. Using Fisher's exact tests to examine the enrichment of microRNA targets per transcript compared with other microRNAs 6 microRNA-transcript combinations were robust to Bonferroni correction for multiple testing listed in Table 4.3.

The top three enriched genes here: SERF2 (Small EDRK-Rich Factor 2) with two binding sites for miR-1282, TARS2 (an aminoacyl-tRNA synthetase) with two binding sites for miR-6878, and TRIP6 (Thyroid Hormone Receptor Interactor 6) with two binding sites for miR-6875, are genes where the listed microRNA is also transcribed from the same locus, and the CLASH binding sites are adjacent to the annotated microRNA loci. In these cases these target sites may be due to un-annotated partner mature microRNAs, where the hybrid is actually a microRNA-microRNA hybrid.

TP-73-AS1 an antisense non-coding RNA has three separate binding sites for miR-125a-5p,

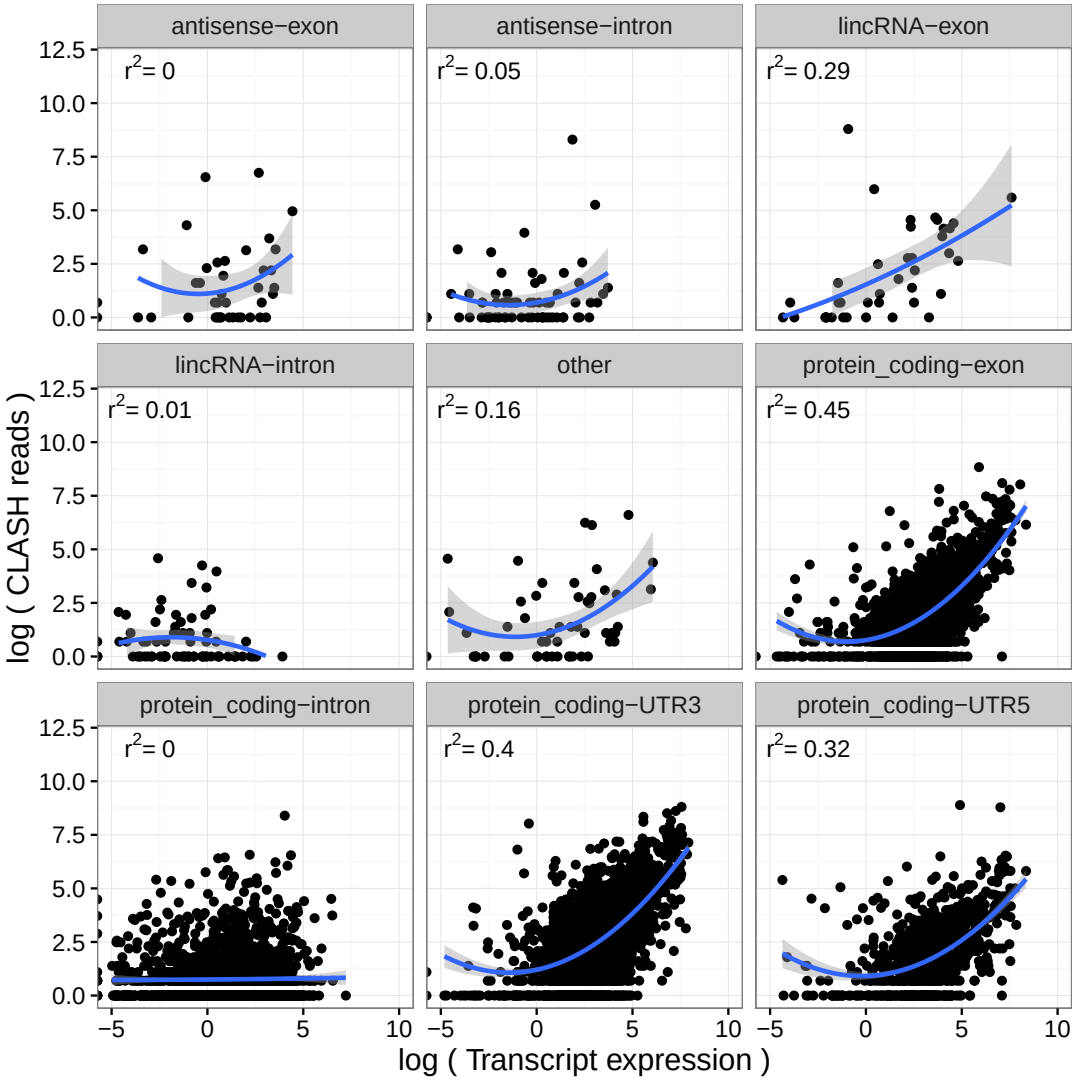
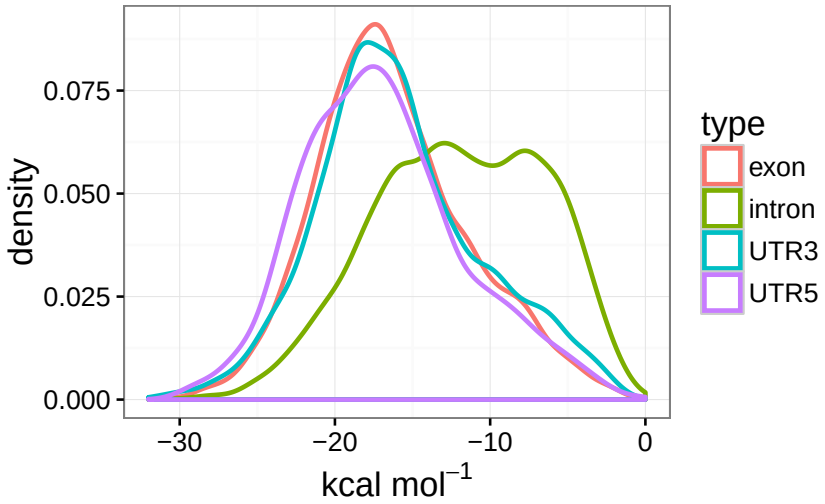
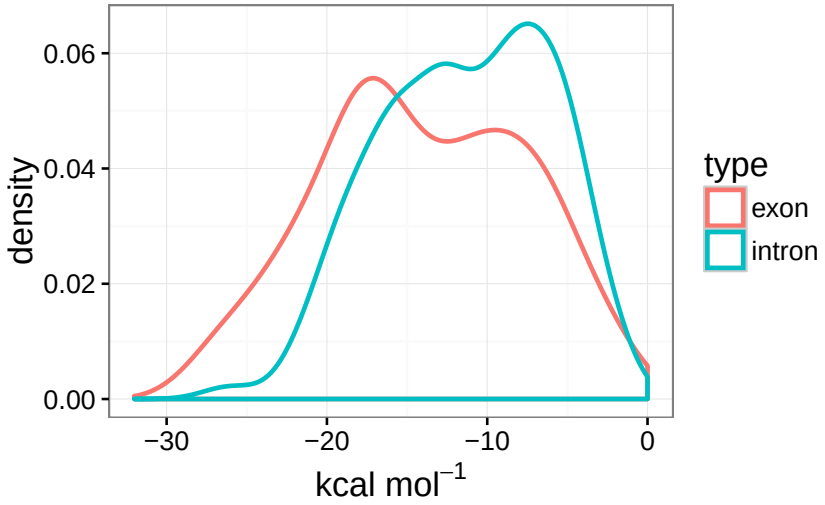


Figure 4.7: A comparison of CLASH and RNA-seq reads assigned to transcripts. (X-axis) log of TPM value for each transcript calculated by Kallisto from whole cell RNAseq. (Y-axis) log of CLASH reads annotated to each transcript. A separate graph is drawn for each class of transcripts, the 'other' class contains other classes of transcripts listed in the Gencode v19[205] annotation: processed transcripts, sense intronic, sense overlapping. Lines of best fit are shown calculated with linear regression using second order polynomials, 95% confidence intervals for this line are shown in grey, adjusted r^2 values for these models are shown. Amongst non-intronic transcript segments CLASH binding is correlated with gene expression.



(a) Protein-coding genes



(b) lincRNAs

Figure 4.8: Folding energy distribution of microRNA:target chimeras in (a) segments of protein-coding genes (b) lincRNAs. More negative kcal/mol values indicate more strongly base-paired interactions.

gene	microRNA	gene miRNA	gene others	non-gene miRNA	non-gene others	pvalue
SERF2	miR-1282	2	0	1	15246	2.58E-08
TARS2	miR-6878-3p	2	2	0	15245	5.16E-08
TRIP6	miR-6875-3p	2	2	0	15245	5.16E-08
TP73-AS1	miR-125a-5p	3	0	86	15160	1.92E-07
MAP3K1	let-7g-5p	2	1	11	15235	2.01E-06
CDR1	miR-7-5p	2	1	21	15225	6.52E-06

Table 4.3: Genes enriched for specific microRNA targets. Numbers represent counts of targets for the microRNA listed or all other microRNAs, within the gene listed or within all other genes. Making the contingency table for a Fisher's exact test pvalue shown.

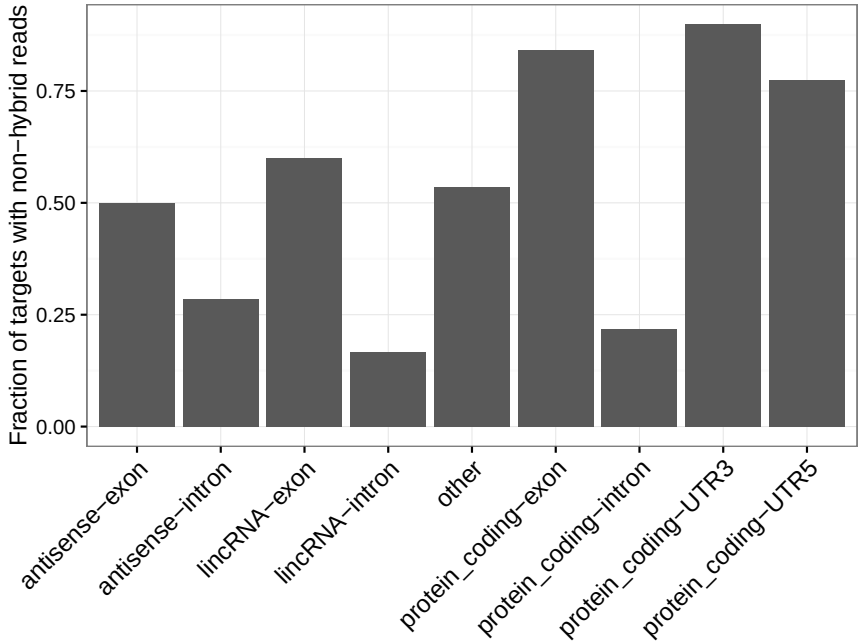


Figure 4.9: The fraction of microRNA targets with supporting non-hybrid reads. (X axis) Transcript annotation category. (Y axis) Fraction of microRNA targets with supporting non-hybrid reads. As hybrid reads are a subset of CLASH reads, additional supporting evidence from non-hybrid reads indicates a more reliable interaction.

with no other microRNA targets. These hybrids have been identified from 1, 1 and 2 sequencing reads with no non-hybrid reads overlapping their target sites. This would be a potential genuine target or competitive RNA for miR-125a.

MAP3K1 has two targets for let-7g-5p in its 3'UTR, the only other microRNA targeting this gene is let-7a with a target site overlapping one of the let-7g sites. This may represent genuine gene target for the let-7 family.

CDR1 (Cerebellar Degeneration Related Protein) with two binding sites for miR-7 is an example of a circular competitive RNA previously identified to compete for miR-7 binding[160], discussed more in section 4.3.7.

Using the number of CLASH reads for a target as a proxy of the strength or reliability of binding and examining each microRNA individually will allow the identification of the classes of transcript they bind and allow the identification of those transcripts which are highly bound by Argonaute compared to their expression level. Fig 4.10 shows scatter plots comparing the number of CLASH reads to transcript expression level for the 12 microRNAs with the highest number of targets. Lowly expressed transcript on these graphs with a high number of CLASH reads are likely transcripts such as short RNAs which are poorly captured by the RNA-seq procedure used. Mitochondrial RNAs shown in green while amongst the most highly expressed transcripts detected are often bound less than would be expected for their expression level, appearing below the lines of best fit.

The linear models shown as the lines of best fit for these 12 microRNAs have adjusted R squared values of between 0.22 and 0.51 (median 0.31), demonstrating that transcript expression level explains a large fraction of the variance in this data.

10 Genes with the largest difference between the linear model predicted value and actual for CLASH reads are labelled, representing candidates for genuine targeting for repression or a competitive endogenous RNA (ceRNA) effect. Several transcripts are commonly found in these lists of genes with more CLASH reads than predicted, they include highly expressed such as ribosomal proteins (RPL30, RPL31) and tubulin (TUBA1B) with TPM>1000 in the top 0.7% of genes by expression. Other genes bound more strongly than expected by expression value by multiple microRNAs are; DHFR, HSPA1A, GLUL, GRAMD1A, VDAC1, REEP4, EIF3CL, and SMARCC1. With TPM ranging between 40 and 200 in the top 20% of genes by expression. These genes may represent those which are poorly captured in the RNA-sequencing experiment, or may be common targets able to bind easily to a variety of microRNAs.

Other genes bound more than predicted for a single microRNA using the linear model based on their expression are listed in Table 4.4. These may be candidates for gene-specific targets for repression or competitive endogenous RNAs (ceRNAs) with single, but highly bound by Argonaute target sites.

4.3.4 Modelling microRNA targeting

As discussed above linear models using only transcript expression as a predictor can explain ∼28% of the variation in CLASH reads per transcript. Here I use additional features of microRNA targeting to model the variation in CLASH reads for each transcript. Using a random forest algorithm I examine the parameters which best predict CLASH reads per transcript. Features included in the model are: transcript expression level (whole cell, cytosolic, or nuclear), GC content of target site, distance from stop codon (for protein coding transcripts), folding energy (a measure of how well the microRNA and its target base-pair), seed type (whether base-pairing is present for 6, 7 or 8 bases of the seed region) and the part of the transcript a microRNA is targeting (intron/exon/5'UTR/3'UTR).

Creating separate random forest models for each of 23 microRNAs which have more than 200 CLASH hybrid loci with these parameters, models have r^2 values between 0.19 and 0.61 (median 0.39) showing a wide variation in how well these models explain the variation in assigned CLASH reads. This suggests that the features predictive of targeting may differ between microRNAs, with some not being captured by the models in use here.

Fig 4.11 shows a boxplot with the distribution in variable importance measures assigned to the each parameter for each of the 23 microRNAs analysed. As expected expression values for either whole cell or a cellular fraction are the most important variables, with whole cell more often the most predictive.

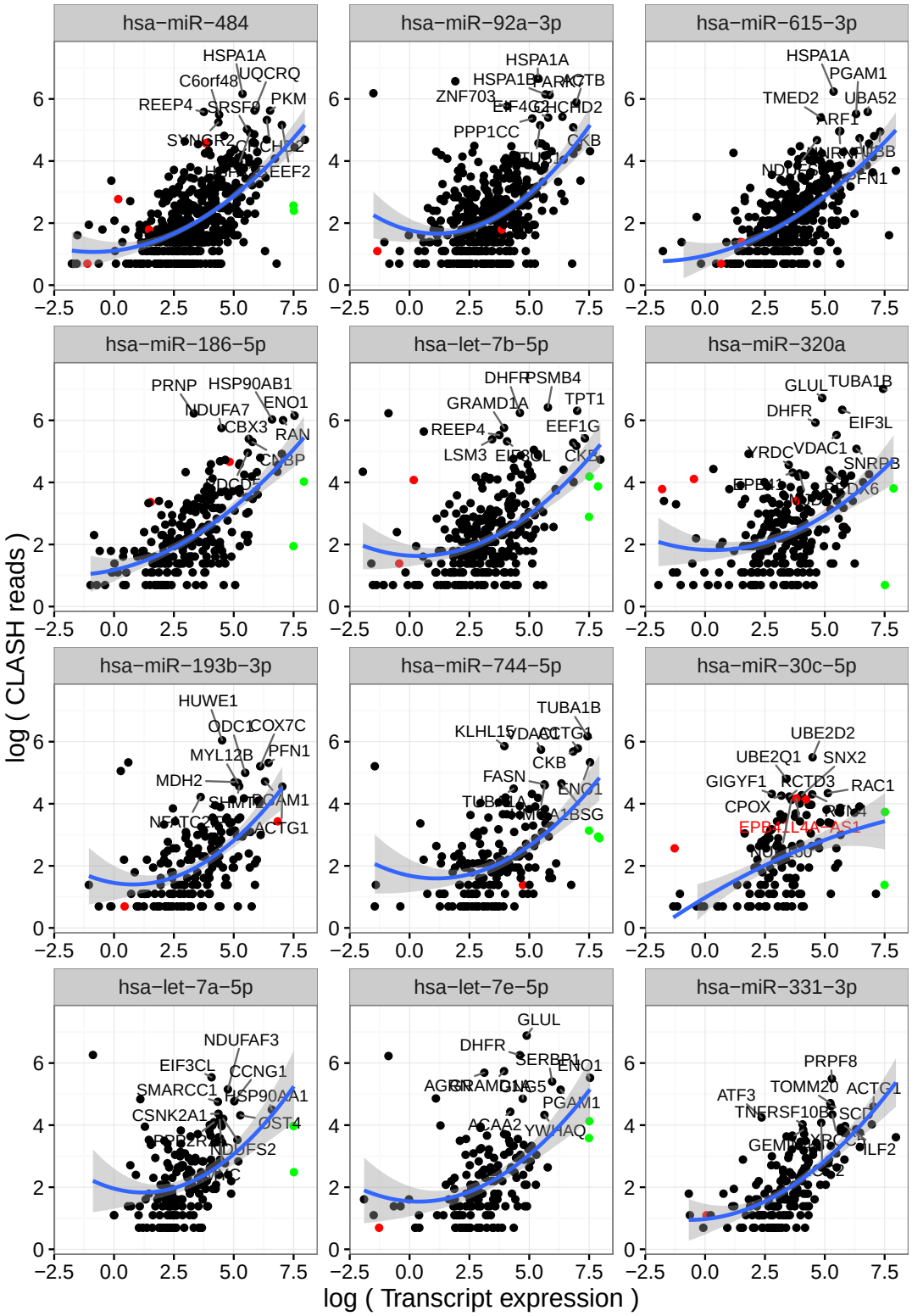


Figure 4.10: Comparison of transcript expression and number of CLASH reads. (X-axis) log of TPM value for each transcript calculated by Kallisto from whole cell RNAseq. (Y-axis) log of CLASH reads annotated to each transcript. A separate graph is drawn for each microRNA. Points are coloured by the transcript class they represent: black=protein-coding, red=lincRNA, green=mitochondrial RNA. Lines of best fit are shown calculated with linear regression using second order polynomials, 95% confidence intervals for this line are shown in grey. The top 10 transcripts ordered by difference between value predicted by the linear model and actual number of CLASH reads for each microRNA with >10 TPM are labelled.

microRNA	gene	expression (TPM)	actual reads	predicted reads
miR-484	UQCRQ	353.6	278	20.7
miR-484	PKM	678.6	277	29.8
miR-484	C6orf48	80.7	242	9.8
miR-484	SYNGR2	78.1	190	9.6
miR-484	SRSF9	258.5	152	17.5
miR-484	H2AFX	99.9	123	10.8
miR-484	COX5A	350.5	129	20.6
miR-92a-3p	PARK7	347.0	458	20.5
miR-92a-3p	ZNF703	59.1	320	8.5
miR-92a-3p	PPP1CC	168.4	215	14.0
miR-92a-3p	STUB1	234.0	173	16.6
miR-615-3p	TMED2	125.4	222	12.1
miR-186-5p	NDUFA7	88.9	315	10.2
miR-186-5p	EEF1A1	4907.3	334	104.3
miR-186-5p	CBX3	278.8	222	18.2
miR-186-5p	CNBP	324.3	201	19.7
miR-186-5p	PDCD5	266.0	142	17.7
let-7b-5p	PSMB4	323.4	612	19.7
let-7b-5p	LSM3	31.4	219	6.4
let-7b-5p	EEF1G	1518.7	226	48.4
miR-320a	EIF3L	307.4	567	19.2
miR-320a	SNRPB	557.2	161	26.6
miR-30c-5p	UBE2D2	89.1	244	10.2
let-7a-5p	NDUFAF3	116.3	173	11.6
let-7e-5p	AGRN	22.7	296	5.6
let-7e-5p	SERBP1	386.0	222	21.7
let-7e-5p	GNG5	113.0	128	11.5

Table 4.4: Genes with $>$100 more CLASH reads than predicted using the linear models displayed in figure 4.10 based on their expression in transcripts per million (TPM).

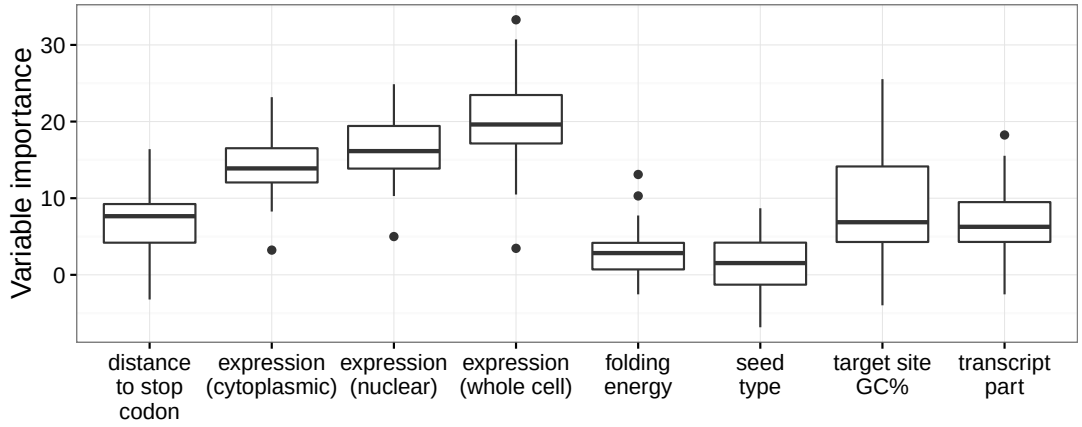


Figure 4.11: Variable importance in predicting CLASH binding. (X-axis) random forest model parameter. (Y-axis) variable importance measure calculated using AUC based variable permutation[234]

Distance to stop codon is significantly predictive for 3'UTR and exon targets with a smaller distance being predictive of higher numbers of reads, suggesting that targets at final exons and at the 5' end of 3'UTRs were stronger hybrids with more CLASH reads.

Transcript part (exon, intron, 5'UTR, 3'UTR) is predictive with intron targets having a negative weight, 3'UTR and exon targets having a positive weight. 5'UTR targets are not significantly predictive. GC content of target site is also shown as a useful predictor here, with higher GC negatively correlated to CLASH reads.

Folding energy and seed type (8mer, 7mer, 6mer) have been previously seen to be a useful factor for predicting the activity of microRNAs[96], and folding energy has been used previously in this chapter as a measure of specificity having different distributions between exonic and intronic targets (Fig 4.8). However in this analysis both are the least predictive parameters for the number of CLASH reads present at a targeted locus. This may be because while a well base-paired microRNA with binding in the seed region are required for target repression they are not necessary for target binding, with perhaps a larger pool of more poorly bound transient targets.

4.3.5 microRNAs targeting transcription start sites

A number of studies have suggested a role for microRNAs in transcriptional gene regulation[137, 231, 139, 132, 144, 147, 141] (section 1.5.3). Here I test this hypothesis by examining for enrichment of specific microRNAs with targets around transcription start sites. 11,385 robust transcription start site loci were obtained in HEK293 cells from cap analysis of gene expression sequencing (CAGE-seq)[236] as part of the FANTOM5 project[207].

Intersecting regions upstream of these TSSs identified 936, 704, and 584 microRNA targets within 200bp, 100bp and 50bp upstream of TSSs respectively.

microRNA	tss miRNA	tss others	non-tss miRNA	non-tss others	pvalue
let-7c-5p	10	339	116	22809	2.1E-05

Table 4.5: MicroRNAs enriched for CLASH targets within 50bp upstream of TSSs. Numbers represent counts of targets for the microRNA listed or all other microRNAs, within 50bp upstream of the TSSs or all other targets. Making the contingency table for a Fisher's exact test, pvalue shown.

microRNA	tss miRNA	tss others	non-tss miRNA	non-tss others	pvalue
miR-1307-3p	9	362	51	22256	5.4E-07
miR-935	9	362	75	22232	9.5E-06

Table 4.6: MicroRNAs enriched for CLASH targets within 50bp downstream of TSSs. Numbers represent counts of targets for the microRNA listed or all other microRNAs, within 50bp downstream of the TSSs or all other targets. Making the contingency table for a Fisher's exact test, pvalue shown.

Testing microRNAs with more than 10 hybrids excluding those targeting repeat sequences and microRNA loci using Fisher's exact tests for enrichment of targets within 50bp,100bp or 200bp upstream of FANTOM5 transcription start sites identified 1 significant microRNA robust to Bonferroni multiple testing correction shown in Table 4.5. Let-7c was significantly enriched for binding upstream of TSSs because of 12 binding sites mapping to RNAs encoded on the mitochondrial chromosome, listed as upstream regions due to the gene density of the region. Other than this there was no enrichment for particular microRNAs binding at transcription start sites.

Intersecting regions downstream of these TSSs identified 1838, 1303, and 939 microRNA targets within 200bp, 100bp and 50bp downstream of TSSs respectively.

Testing was performed microRNAs with more than 10 hybrids excluding those targeting repeat sequences and microRNA loci using Fisher's exact tests for enrichment of targets within 50bp,100bp or 200bp downstream of FANTOM5 transcription start sites. This identified two microRNAs listed in Table 4.6: miR-1307-3p and miR-935 each with 9 targets within 50bp downstream of TSSs. These targets seemed to be biased towards highly expressed genes with the median TPM for miR-1307-5p TSS targets being 369 TPM in the top 1% of transcripts by expression. For miR-935 the median TPM for TSS targets was 33 TPM in the top 10% of transcripts by expression.

4.3.6 microRNAs targeting splice sites

A number of studies have suggested a role for Argonaute in altering splicing[154, 153, 151, 150] (section 1.5.4), here I examine the distribution of microRNA targets around splice sites and test for enrichment of specific microRNAs with targets around splice sites.

Intersecting the 3p and 5p ends of introns from Gencode v19[205] with CLASH microRNA

microRNA	splice miRNA	splice others	non-splice miRNA	non-splice others	pvalue
miR-615-3p	345	6109	518	15706	8.3E-14
miR-92a-3p	377	6077	621	15603	5.0E-11
miR-149-5p	136	6318	203	16021	2.2E-06
miR-92b-3p	79	6375	104	16120	1.4E-05
miR-222-3p	117	6337	180	16044	2.9E-05
miR-103a-3p	127	6327	211	16013	1.6E-04
miR-484	364	6090	732	15492	2.4E-04

Table 4.7: MicroRNAs enriched for CLASH targets within 50bp of splice sites. Numbers represent counts of targets for the microRNA listed or all other microRNAs, within 50bp of splice sites or all other targets. Making the contingency table for a Fisher's exact test, pvalue shown.

targets identified 7092, 9256, and 11375 targets allowing windows of 50p, 100p, and 200bp respectively, representing 30%, 40%, and 49% of protein-coding and lincRNA targets due to the density of targets in exonic sequence.

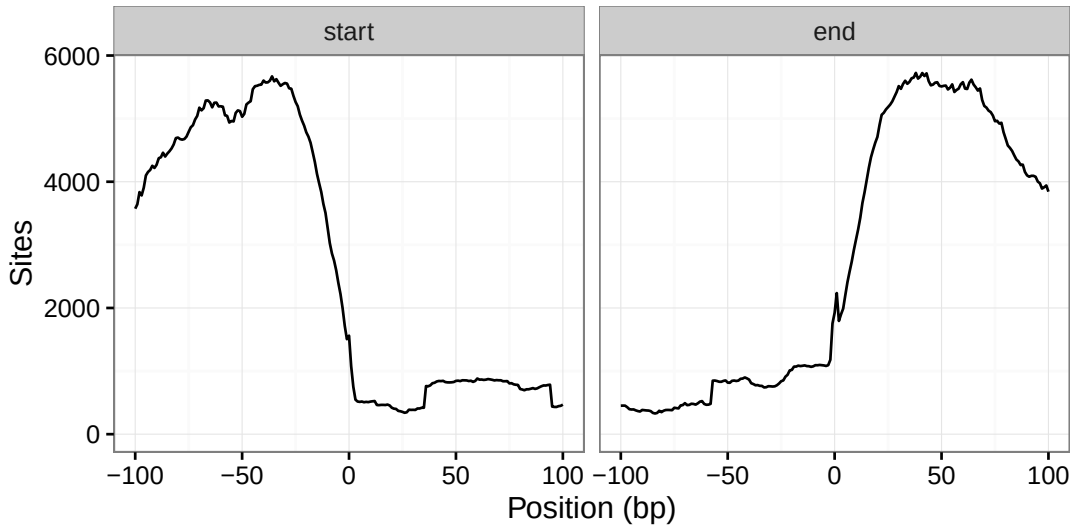
Testing microRNAs with more than 10 hybrids excluding those targeting repeat sequences and microRNA loci using Fisher's exact tests for enrichment of targets within 50bp of splice junctions identified 7 microRNAs listed in Table 4.7.

100bp flanking either side of Gencode v19[205] introns were intersected with microRNA target sites to examine the distribution of targets around splice sites. Fig 4.12a shows that targets were more highly enriched within the exons. However examining individual microRNAs showed a peak around the end of the intron at splice acceptor sites for microRNA miR-320a seen in Fig 4.12b. Although miR-320a was not enriched at splice sites compared to other targets in the Fisher's exact tests above with 210 target sites within 10bp of splice junctions, and 330 other targets giving an odds ratio of 1.01 and pvalue 0.15 compared with all other microRNAs.

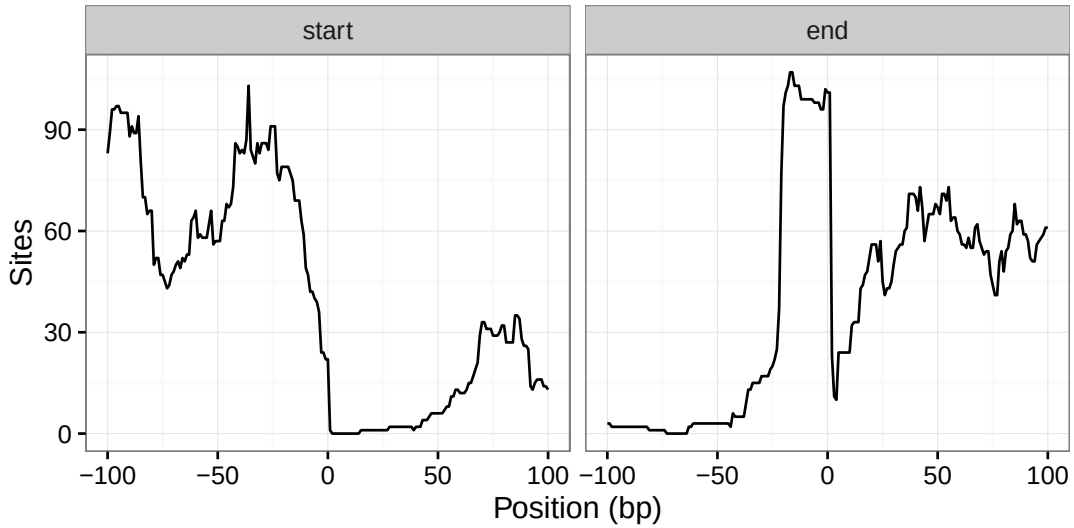
4.3.7 microRNAs targeting circular RNA 'sponges'

Circular RNAs have been observed in human cells and are suggested to act as ceRNAs or 'microRNA sponges' acting to sequester microRNAs[159, 160, 161] (section 1.6). Intersecting CLASH microRNA hybrids with a dataset of 1953 circular RNAs identified in HEK293 cells by Memczak et al[160] I find 602 circRNA-microRNA pairs involving 322 circRNAs and 162 microRNAs.

Using Fisher's exact tests to identify microRNAs which were enriched for targeting predicted circular RNAs, compared with all other microRNA targets, no microRNAs were robust to Bonferroni correction. Those passing nominal significance are shown in Table 4.8, the most significant of these miR-7 with three targets in predicted circular RNAs was the microRNA identified in Memczak et al.[160] as being sequestered by a circular RNA named CDR1as. Two of the three binding sites listed for this microRNA in Table 4.8 were within CDR1as.



(a) All microRNAs



(b) microRNA miR-320a

Figure 4.12: Distributions of microRNA targets around spice junctions. (Left panels) 200bp around splice donor sites. (Right panels) 200bp around splice acceptor sites. (A) All microRNA target sites. (B) miR-320a target sites.

microRNA	circ miRNA	circ others	non-circ miRNA	non-circ others	pvalue
miR-7-5p	3	360	34	22281	0.021
miR-92b-3p	7	356	176	22139	0.028
let-7g-5p	2	361	18	22297	0.040
miR-125a-3p	2	361	18	22297	0.040
miR-6807-3p	3	360	47	22268	0.046

Table 4.8: MicroRNAs enriched for CLASH targets within circular RNAs. Numbers represent counts of targets for the microRNA listed or all other microRNAs, within predicted circular RNAs or all other targets. Making the contingency table for a Fisher's exact test, pvalue shown.

4.4 Chapter summary

Applying the genome-wide hybrid mapping pipeline discussed in the previous chapter to nine AGO1 CLASH datasets from Helwak et al. identified ∼40,000 microRNA binding sites across a variety of transcript classes. Most abundant amongst these were targets within protein coding genes, particularly 3'UTRs as expected for microRNA mediated gene repression, and also within internal exons, 5'UTRs and introns. Another abundant class of targets were repetitive elements, particularly rRNAs, tRNAs, LINEs, and SINEs. Also present were targets within lincRNAs, pseudogenes, and antisense RNAs. These microRNA targets across the length of protein-coding transcripts and in a variety of non-protein-coding transcripts may represent chance binding events due to partial complementarity or may be functional for an effect in controlling free microRNA levels or other affects mediated by Argonaute interactions with other effector proteins.

A number of studies have suggested a role for microRNAs in affecting transcription where being targeted near TSSs Argonaute can interact with chromatin remodellers to affect accessibility. Testing for enrichment of microRNA targeting upstream of TSSs revealed only an artefact due to a high level of Let-7c binding to mitochondrial genes which have many close TSSs in the CAGE dataset. Two microRNAs miR-1307-5p and miR-935 were enriched for targeting downstream of proximal TSSs, each targeting nine genes within 50bp of a TSS, biased towards highly expressed genes.

Some studies have also suggested a role for microRNAs in affecting splicing via Argonaute interactions with splicing factors. Testing microRNAs for enrichment of targets around splicing acceptor or donor sites seven microRNAs were enriched for targets within 50bp, although 30% of protein coding or lincRNA targets fell within these windows and each of these seven microRNAs had more non-splice-site targets than splice-site targets suggesting these interactions were not specific. Similarly the peak of miR-320a binding around splicing acceptor sites is likely due to complementarity to the consensus sequence around this site, despite this miR-320a was not enriched for binding at these regions compared to other microRNAs.

Circular RNAs have been hypothesised as a source of ceRNAs with one striking example for miR-7. No microRNAs were significantly associated with predicted circular RNAs, although the microRNA with the most significant pvalue was for miR-7 with its previously identified ceRNA CDR1as. This result would support the hypothesis that the majority of circular RNAs were low abundance splice variants.

Testing amongst all Gencode v19[205] transcripts for those enriched with multiple binding sites identified: The CDR1 transcript previously known with miR-7 binding sites. MAP3K1 with binding sites for let-7a/let-7g and TP73 antisense transcript with 3 binding sites for miR-125a-5p candidate targets or ceRNAs. And three transcripts SERF1, TARS2 and TRIP6 with multiple binding sites for microRNAs where the pre-microRNAs are contained within these genes, probably representing their mis-annotated or un-annotated mature microRNAs.

Examining correlations between transcript abundance in RNA-seq data and the level of CLASH binding to transcripts positive correlations could be seen above a transcript abundance threshold for exon lincRNA targets and exon, 3'UTR, and 5'UTR protein-coding gene targets. These correlations were not seen for intronic targets and were less strong in antisense RNAs. It would be expected that transcript expression level would strongly correlate with microRNA targeting, where these correlations are not present for intronic targeting it may be due to transcripts which are not well captured in this RNA-seq dataset. Additionally candidates for targets for repression or ceRNAs may be seen as deviations from this expected level of targeting given transcript abundance, some of which appear to be common to several of the microRNAs examined while others are microRNA specific.

Weaker or more transient targeting in introns was also suggested by the calculated RNA folding energy as a measure of how well microRNAs base-paired with their targets which was weaker for targets at intronic sites, and similarly strong in protein-coding exons, 3'UTRs, 5'UTRs and lincRNA exons.

Modelling the correlation between expression level and CLASH targeting on a per-microRNA basis revealed differences in the strength of this correlation. Examining those transcripts which are bound by specific microRNAs more often than would be expected for their expression level demonstrated candidate targets for repression and potential competitive RNAs.

Modelling other parameters which may contribute to microRNA targeting revealed that while target expression was most predictive, also predictive were; smaller distance to stop codon, exon or 3'UTR targets and target site GC%. Folding energy and binding in the seed region, while previously seen as important for transcript repression were not predictive for the level of CLASH reads at a target site, perhaps due to a larger pool of more weakly bound but not repressed targets being analysed here.

As well as deviations from expected levels of CLASH binding according to transcript abundance, a clustering approach may identify transcripts or regions of transcripts which were more highly bound by Argonaute than expected by chance. This will be examined in the next chapter.

Chapter 5

Clustering of microRNA targets

5.1 Introduction

microRNAs as part of the RISC complex repress their protein-coding targets through complementary base-pairing with sites in their 3'UTR (section 1.5). This activity is dependent on the microRNA:target ratio, where additional microRNA targets above a certain threshold will act to reduce the repression on microRNA targets. It has been hypothesised that in order to affect microRNA repression a substantial number of additional target sites will be necessary requiring highly expressed transcripts with small numbers of binding sites or expression of transcripts with a large number of microRNA target sites (section 1.6).

Chapter 4 examined transcripts for the enrichment of microRNA hybrids involving the same microRNA, identifying three transcripts TP73-AS1 and miR-125a, maP3K1 and let-7, and CDR1 and miR-7.

This chapter examines the spatial clustering of all microRNA hybrids within the genome to identify exons or regions of genes which are densely bound by Argonaute, and may represent candidate competitive endogenous RNAs (ceRNAs).

5.2 Methods

5.2.1 Data

Datasets analysed here are the CLASH hybrid and CLASH non-hybrid genomic locations from chapter 4. These datasets were separately merged using the BedTools merge function such that several hybrids targeting the same site would be counted as a single site for clustering.

5.2.2 Genome-wide clustering

A clustering approach was implemented in R to identify spatial clusters of binding sites using a dynamic programming approach. This algorithm generated scores along the genome such that:

- A new binding sites added a score of **m**
- Within a binding site scores remained the same
- Not within a binding site scores decay to 0 by **n** per base.

Backtracking is then performed to identify clusters in order of their maximum score above a threshold:

- Starting with the highest score, the region before that point until the score reaches 0 is recorded as a cluster
- That region is masked
- Scores for the region beyond that maximum score point are recalculated, and the next highest score is recorded

This approach allows clusters to grow dynamically rather than using fixed size windows. Selected parameters for the algorithm were a score of 30 for each new binding site with a decay of 0.2 per base, clusters were identified with a score above 31. Sample results of the clustering algorithm are shown in Fig 5.1.

5.3 Results

5.4 Clustering microRNA binding sites

Applying the clustering algorithm to CLASH microRNA targets identified 349 cluster of multiple target sites. Counts of these clusters with the Gencode v19[205] transcripts they overlap with are shown in Fig 5.2, often clusters would overlap multiple regions of protein-coding genes due to alternatively spliced transcripts or alternative 3'UTR sites.

Only six clusters of microRNA targets did not overlap any Gencode v19[205] transcript (table 5.1). Two of these (IDs 27 and 249) corresponded to small nucleolar RNAs not present in Gencode v19. The remaining four regions had overlapping mRNA sequences from GenBank seen on the UCSC genome browser, but no annotated transcripts.

Eleven clusters of microRNA targets are present within predicted circular RNAs by Memczak et al.[160]. Shown in table 5.2 these results are biased towards highly expressed genes providing a greater opportunity for base-pairing in a small genomic region. Clusters with the MYH9, KPNB1, DDX5, HNRNPA2B1 and CANX genes are present at predicted circularised

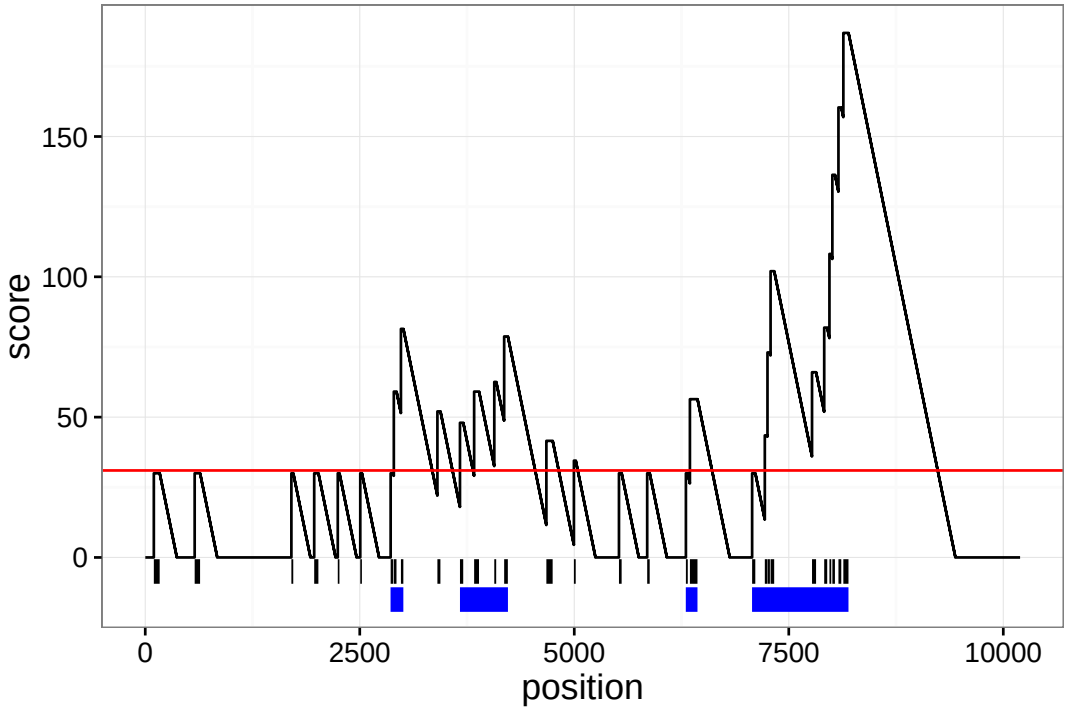


Figure 5.1: Example scores generated by the clustering method. Black boxes below 0 represent sites of Argonaute binding, the score calculated in the first pass is shown on the Y axis, new binding sites increase the score while there is a drop-off penalty. The red line shows the threshold for clusters with more than one binding site. Clusters identified by the algorithm described in the text are shown as blue segments.

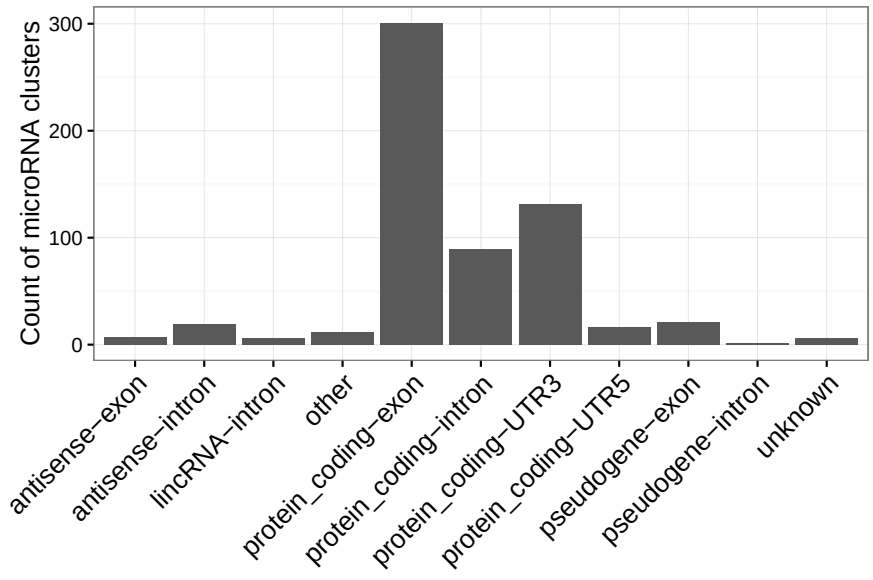


Figure 5.2: The count of microRNA hybrid clusters from hybrid CLASH reads annotated to different transcript classes. (X axis) Transcript annotation. (Y axis) Count of microRNA clusters annotated to that class.

ID	chr	start	end	length	score
27	chr19	50595748	50595865	117	58
41	chr13	110076545	110076633	88	57
57	chrM	240	411	171	57
86	chr2	71127445	71127491	46	55
222	chr7	10492980	10493044	64	46
249	chr19	48421688	48421804	116	42

Table 5.1: Clusters of microRNA targets sites with no gencode v19 annotation corresponding to unannotated transcripts

gene	region	chr	start	end	score	tpm
MYH9	exon	chr22	36684297	36684391	47	50.6
KPNB1	exon	chr17	45747066	45747127	55	334.5
DDX5	exon	chr17	62500843	62500905	57	359.7
BCL7A	intron	chr12	122473442	122473506	32	19.4
HNRNPA2B1	exon	chr7	26235465	26235531	52	1258.8
CANX	exon	chr5	179132682	179132731	58	247.4
HIST1H1E	exon	chr6	26156943	26156997	56	0.3
HIST1H2AG	exon	chr6	27101195	27101287	41	0.7
HIST1H2BK	exon	chr6	27114559	27114616	56	35.9
HIST1H2BK	exon	chr6	27114152	27114273	46	35.9
HIST1H4E	exon	chr6	26205084	26205136	53	0.1

Table 5.2: Clusters of microRNA targets sites within predicted circular RNAs. Score = cluster score, TPM=Transcripts per million (transcript abundance)

exons, or segments of exons. The cluster within an intron of BCL7A is part of a predicted 13kb circular RNA formed through a retained intron. Five clusters of microRNA binding sites within two histone gene clusters on chromosome 6 are also seen, where circular RNAs are predicted as forming between histone genes. These circular RNAs may be mis-annotations due to similarity of histone genes or genuine trans-splicing events. Transcripts from these genes while not well captured in the RNA-seq experiment will be highly expressed.

Two miR-7 targets within CDR1-as, the previously seen ceRNA, are hybrids located $\sim$1kb apart in this dataset so are not rediscovered as a cluster in this analysis.

5.4.1 Clustering Argonaute binding sites

As captured CLASH microRNA hybrids are likely a subset of sites, examining all Argonaute binding sites through clustering of the non-hybrid CLASH reads may provide a distribution less biased to highly expressed transcripts.

Performing the same clustering approach for these reads finds clusters amongst the transcript types shown in Fig 5.3

Intersecting these hybrids with predicted circular RNAs from Memczak et al.[160] identifies

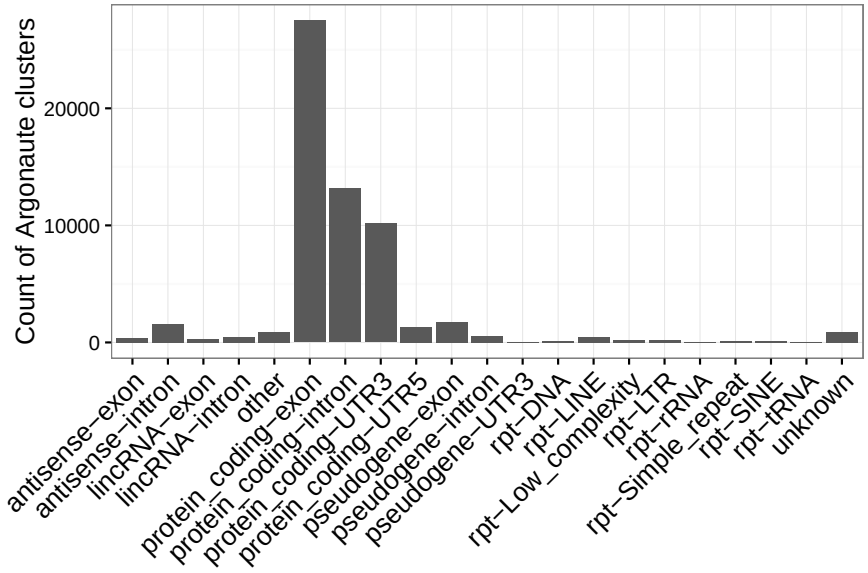


Figure 5.3: The count of Argonaute hybrid clusters identified from all CLASH reads annotated to different transcript classes. (X axis) Transcript annotation. (Y axis) Count of microRNA clusters annotated to that transcript class.

gene	region	chr	start	end	score	tpm
AMOTL1	UTR3	chr11	94604690	94609833	422	12.4
FBXO34	UTR3	chr14	55817619	55820331	338	10.8
MED1	UTR3	chr17	37560733	37566977	336	42.7
MRPS25	UTR3	chr3	15083925	15090778	325	72.7
WDR6	UTR3	chr3	49048976	49051728	320	131.6
MALAT1	exon	chr11	65266354	65273646	317	4.6
RBM12	UTR3	chr20	34237731	34243284	306	54.2
UCK2	exon	chr1	165877971	165880925	286	
LMTK2	exon	chr7	97820924	97823790	281	5.4
ARHGAP35	exon	chr19	47422854	47425055	242	8.5
ATXN7L3B	UTR3	chr12	74931494	74935221	240	48.9
FAT1	exon	chr4	187627716	187630946	224	21.2
LRRC58	UTR3	chr3	120046869	120050077	219	13.6
TMEM194A	UTR3	chr12	57449430	57453844	216	19.8
ZNF544	UTR3	chr19	58772908	58774575	206	34.6
MAP4	UTR3	chr3	47892828	47894621	195	61.0
VKORC1L1	UTR3	chr7	65419060	65422179	191	15.7
SPIN1	UTR3	chr9	91090016	91092406	185	21.1
LMNB2	UTR3	chr19	2428163	2430954	180	65.9
DYNLL2	UTR3	chr17	56166500	56169397	179	20.5

Table 5.3: Clusters of Argonaute binding sites within predicted circular RNAs. Score = cluster score, TPM=Transcripts per million (transcript abundance)

459 clusters amongst 364 predicted circular RNAs. The majority of these are annotated as either protein-coding exons, or protein-coding 3'UTRs. The top 20 of these clusters by score are listed in table5.3, where a cluster within the lincRNA MALAT1 is also seen.

5.5 Chapter summary

In chapter 4 the enrichment for the binding of specific microRNAs to transcripts was examined. In this chapter multiple microRNA target sites or Argonaute binding sites within relatively small regions of the genome were analysed regardless of identified targeting small RNAs. These clusters of binding sites were typically present at protein-coding UTRs and protein coding exons.

3'UTRs are an expected target of genuine transcript repression, whereas circularised protein-coding exons have been hypothesised as a source of competitive endogenous RNAs. Few clusters of binding sites were present outside of Gencode v19[205].

It was also observed in chapter 4 that gene expression was a major factor in observing microRNA binding, and clusters of microRNA targets sites seen here are biased towards highly expressed transcripts.

Examining clusters of Argonaute binding from non-hybrid reads, a more dense dataset, identifies spatial clusters seen here often within single exons or 3'UTRs which are less biased towards highly expressed transcripts, some of which are present within circular RNAs and may represent good candidates for competitive endogenous RNAs.

As few clusters of binding sites in unexpected locations were identified, and few enrichments for the binding of specific microRNAs to transcripts was seen these results have not been followed up in detail.

The remaining chapters of this thesis explore the second of my project aims: to examine the role of polymorphisms at microRNA loci in microRNA processing, and effects in trait and disease phenotypes.

Chapter 6

MicroRNA processing quantitative trait variants (mpQTVs)

6.1 Introduction

The key pathways in the production of microRNAs in humans have been uncovered with further details of their mechanisms continuing to be elucidated (section 1.3.1). In the canonical microRNA biogenesis pathway microRNAs are transcribed as pri-microRNAs containing one or more regions which form hairpin loops in their secondary structure. These hairpin loops are recognised and cut in the nucleus to form pre-microRNAs by the microprocessor complex comprised of the double stranded RNA binding protein DGCR8 and the RNase III enzyme dicer. pre-microRNAs are then exported to the cytoplasm where a second processing event is carried out by another RNase III enzyme Dicer producing mature microRNAs which may then be loaded into Argonaute proteins to form the RISC complex which can regulate the expression of other genes.

The effect of genetic variation between individuals in affecting microRNA biogenesis has been examined in a number of studies correlating gene expression measures from expression arrays or small RNA sequencing with genotype data from SNP genotyping arrays or whole-genome sequencing to identify eQTLs (section 1.7.2).

As the power of these studies has increased through sample size and density of genotyping they have identified increasing numbers of microRNA eQTLs. Using SNP genotyping arrays finding cis-eQTLs for 6[183], 12[182], and 14[184] microRNAs. And using samples whole-genome sequenced through the 1000 Genomes Project has identified cis-eQTLs for 31[185] and 60[186] microRNAs. The most recent and largest of these studies was performed by the GEUVADIS consortium[186] utilising small RNA sequencing in 452 lymphoblastoid cell lines (LCLs) with whole-genome sequencing to an average depth of 5x and exome sequencing to an average depth of 80x from the 1000 Genomes Project[204]. They identified 3868

SNPs which were cis-eQTLs to 60 microRNAs (within 500 kb). However no attempt was made to consider potential functional variants at these quantitative trait loci or identify likely mechanisms explaining the expression-level molecular-phenotypes. Indeed the mechanisms by which eQTL variants affect microRNA expression has only been examined in a handful of cases with disease relevance(section 1.7.4).

In this chapter I examine the distribution of SNP variants at microRNA loci and the evidence for their selective constraint. I examine microRNA cis-eQTL SNPs from the GEUVADIS consortium[186] and identify additional rare cis-eQTL variants, assessing their potential to affect processing. These cis-eQTLs SNPs are then be compared with those that do not affect microRNA expression to examine purifying selection and the features of microRNA hairpins which when disrupted are likely to affect expression leading to a phenotypic consequence.

Identified cis-eQTLs are then prioritised for further analysis such as that applied to the variant in the terminal loop of miR-30c[199] (section 1.7.4) to confirm their effect and examine the mechanism through which they act. I then apply this knowledge about how variants affect microRNA processing to genomic sequencing data from patient cohorts in chapter 7.

6.2 Methods

6.2.1 eQTL Analysis

Common SNPs

The GEUVADIS consortium[186] identified cis-eQTLs from SNPs within 500kb of their curated list of microRNA loci, this list of microRNA loci was an extension of miRBase v18[11] where if only one mature microRNA was annotated from a microRNA hairpin the partner mature microRNA was annotated using RNA structure prediction. Only those mature microRNAs detected as expressed in $>50\%$ of samples were used in eQTL detection, expression values corresponding to read count normalised to the median number of well-mapped reads to account for variation in sequencing depth. Where variants overlapped mature microRNAs sequences the non-reference alleles were generated to account for allelic mapping bias. cis-eQTLs were calculated using the matrixEQTL R package v2.1.1[224].

I have filtered these significant microRNA cis-eQTLs published by the GEUVADIS consortium for variants within 20bp of the miRBase microRNA hairpin they are an eQTL for. This identified 20 SNPs which were cis-eQTL to the microRNA hairpin they were contained within or were adjacent to.

These candidates were then examined as to whether they were previously known or belong to blocks of linkage disequilibrium associated with expression, whether they were the strongest cis-eQTL signal for the microRNA, and whether they disrupted any known structural or

sequence motifs.

Rare SNPs

All bi-allelic SNPs within miRBase v18 microRNA hairpins +/- 20nt were extracted from GEUVADIS vcf files (EBI ArrayExpress accession: E-GEUV-1) using `bcftools`[219] v1.2-5 for testing with the R package `MatrixeQTL`[224] v2.1.1, using an additive linear regression model with a t-statistic test corrected for multiple testing using the Benjamini-Hochberg procedure.

As the power to call significant rare variant eQTLs is reduced relative to common SNPs due to the smaller numbers of rare genotype samples, results were selected based on a less stringent p-value threshold of corrected p-value$<$0.1 as well a threshold on the magnitude of the expression change with the heterozygous genotype having a median expression greater than 0.5 times the standard deviation of the common homozygote away from the common homozygote median expression.

Multivariate linear regression and analysis of variance using the R functions `lm` and `anova` from the stats package with default parameters were used to examine the independence of eQTL signals where multiple SNPs at a loci displayed significance.

6.2.2 Annotation

SNPs were annotated with their microRNA locus finding overlaps for their genomic coordinates with `bedtools intersect`. Based on miRBase annotation of microRNA processing cleavage sites and stem-loop base pairing structure, each microRNA hairpin was split into consecutive segments: 3p and 5p flank, 3p and 5p stem, 3p and 5p mature microRNA, and hairpin loop. SNPs were then annotated with their microRNA region accounting for microRNA strand orientation.

6.2.3 Derived allele frequency tests

The frequency distribution of derived alleles can be used to compare selection at different genomic regions. Derived allele frequency tests were performed using SNPs found in the uniform whole-genome coverage (median 20x) of 2636 individuals in the Icelandic population[210]. Variants were resolved into ancestral and derived alleles through reference to the Ensembl human ancestor reconstructed sequence based on the 12-way mammalian EPO alignments. Variants with unresolved ancestral states and non-single-nucleotide polymorphisms were discarded. The genome-wide allele frequency distribution was used as a proxy for neutral evolution. Where purifying selection acts to remove deleterious alleles an excess of rare alleles will be seen relative to common alleles. Where diversifying selection

acts to increase the allele frequency of new beneficial alleles an excess of common alleles will be seen relative to rare alleles. Statistical tests were performed with a Fisher's exact test using the R function `fisher.text` from the stats package giving the odds ratios and 95% confidence intervals. Common alleles are defined as those with a MAF>5%, rare alleles those with a MAF<1.5%, these values were chosen to maximise the odds ratio while minimising the confidence interval for a comparison between second codon positions and four-fold degenerate sites in amino acid sequences[237].

6.2.4 Odds ratio tests

The odds ratio for SNPs in each microRNA region being a cis-eQTL was calculated with a Fisher's exact test using the R function `fisher.test` from the stats package compared to all SNPs flanking hairpins up to 1kb where ∼3% (196/6431) were cis-eQTLs (corrected pvalue<0.05).

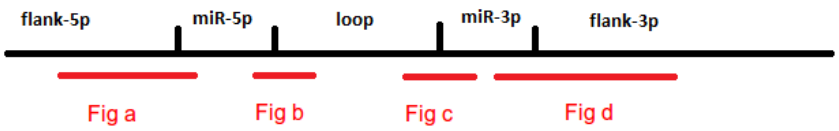
6.3 Results

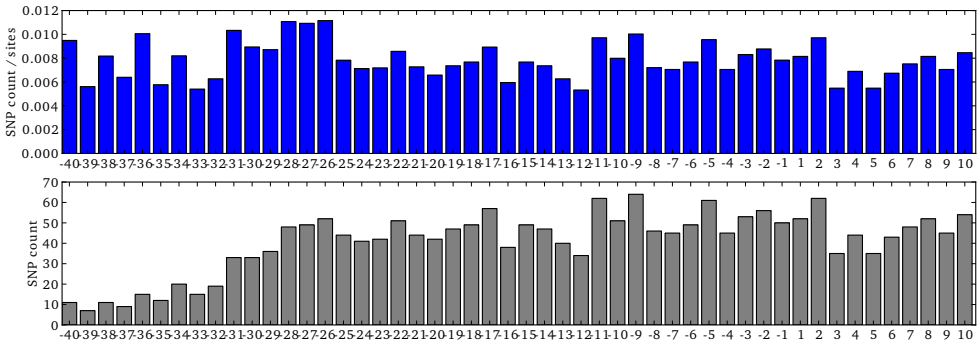
6.3.1 The distribution of common microRNA SNPs

Examining the distribution of SNPs across microRNA loci using all variants found in dbSNP[208] v139 shown in Fig 6.1 we can see many variants are present, fairly evenly distributed across the whole of microRNA loci. A reduced number of common (>5% minor allele frequency) variants may be present around a number of sites: 1bp upstream from the 3p microRNA, and around the base of hairpins at positions -12 to -14 in the 5p-flank and 9 to 11 in the 3p-flank. However this analysis has very little power to detect these changes, and does not account for nucleotide composition.

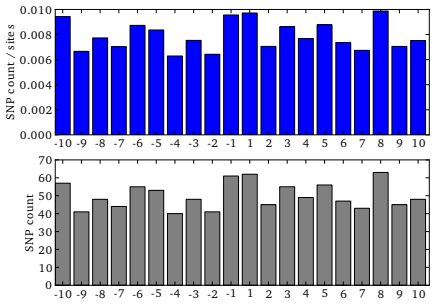
6.3.2 Selective constraint at microRNA loci

To examine the evidence for selection acting at microRNA loci derived allele frequency tests were performed, comparing the frequency of derived alleles across the whole genome as a proxy for neutral evolution to the frequency of derived alleles at whole microRNA loci, and within each annotated microRNA region. Fig 6.2 shows evidence for selective constraint at microRNA loci taken as a whole, although less than at protein coding codons, while in the segment analysis the lower numbers of SNPs give large error bars overlapping the background distribution in some cases.

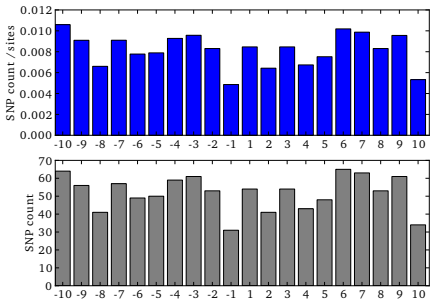




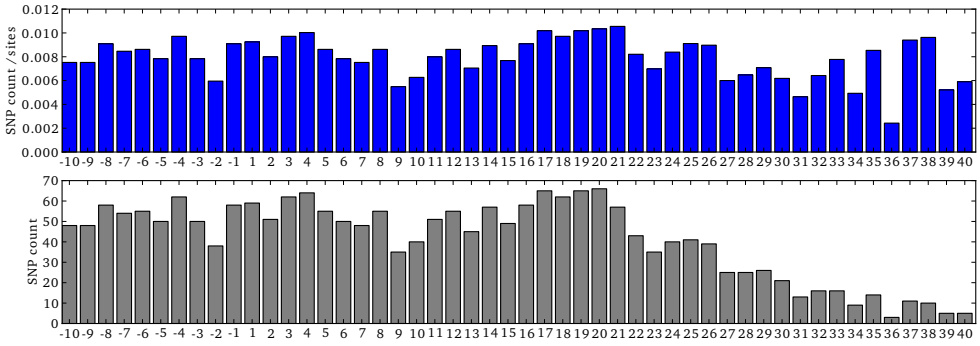
(a) Position around 5p end of miRNA-5p



(b) Position around 3p end of miRNA-5p



(c) Position around 5p end of miRNA-3p



(d) Position around 3p end of miRNA-3p

Figure 6.1: The distribution of dbSNP SNPs at microRNA loci, centred around mature microRNA boundaries shown in the top figure. Grey lower panels: count of SNPs at sites, Blue upper panels: count normalised to the number of sites present.

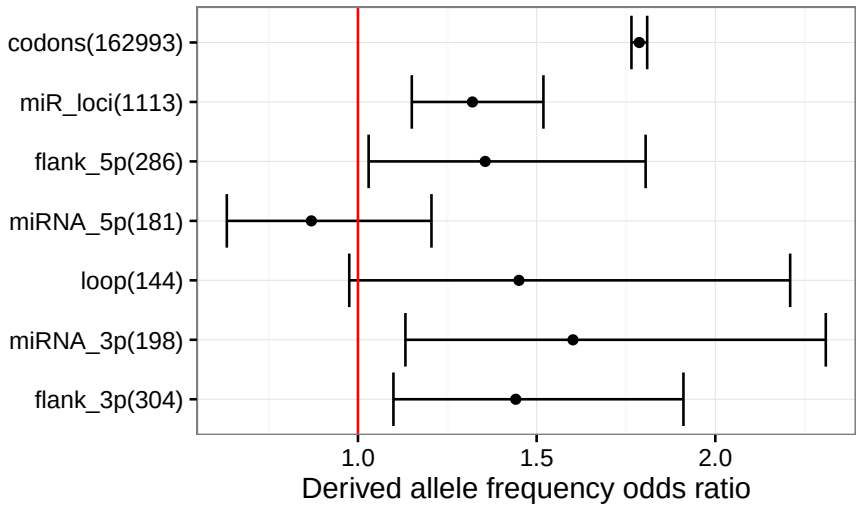


Figure 6.2: Odds ratios of derived allele frequencies for rare(<1.5%) and common(>5%) derived alleles comparing the genome wide distribution to the region labelled, either whole microRNA loci or their annotated segments, protein coding codons are included for comparison. Odds ratios of 1 indicate equality with the genome-wide distribution, higher values indicate selective constraint, and values less than 1 are indicative of net positive selection. The number of SNPs in each category are shown in parenthesis next to the axis labels.

6.3.3 Common microRNA SNPs are cis-eQTLs

Intersecting those common SNPs identified as microRNA eQTLs by the GEUVADIS consortium with miRBase v18[11] precursors +/- 20bp identified 20 SNPs which were a cis-eQTL to a mature microRNA from the hairpin that they were within or adjacent to (Table 6.1). Fig 6.3 shows regional association plots displaying all of the cis-eQTLs up to 500kb away for the selected microRNAs and boxplots displaying the microRNA expression for each genotype of the selected SNP. In some cases the selected SNP present at the microRNA hairpin is the only significant or most significantly associated SNP in the region (miR-3615-3p, miR-3176-3p, miR-4423-5p, miR-4707-3p). In these cases we can infer with reasonable confidence that the selected SNP is likely to be the causal SNP affecting expression of the mature microRNA as with whole-genome sequencing data we have near complete ascertainment of variation. Two related unobserved types of variation could also be in LD with variants selected here; indels often called with lower accuracy may be unobserved or under-observed in this data, and larger segmental duplications/deletions may not have been detected e.g. miR-941 is known to be a multi-copy tandem repeat[91].

Most often the selected intra-hairpin-SNP is amongst a group of SNPs which show similar association as they are contained within a linkage disequilibrium block where a group of SNPs are segregating together (miR-146a-3p, miR-1908-5p, miR-1304-3p, miR-4513-5p, miR-3130-2-5p, miR-1255a-5p, miR-641-5p, miR-1307-3p, miR-92a-1-5p). In these cases it is difficult to predict based on genetic evidence alone which SNP is likely to be the causal SNP, however

the selected SNPs within the microRNA hairpin have the opportunity to affect expression at more stages as they may also directly affect the processing of the primary and precursor microRNA. This hypothesis is being tested experimentally by our collaborators in microRNA processing assays. And in other cases the SNP selected is not the most significant cis-eQTL for the microRNA (miR-618-5p, miR-4745-5p, miR-914-3-3p, miR-5680-3p, miR-4482-1-3p, miR-4741-3p, miR-3188-3p). In these cases the selected SNP may be in weaker LD with a causal SNP flanking the hairpin, or these SNPs may have independent effects on expression. Examining the LD between these SNPs could help to determine the independence of their effects and again processing assays could determine if the selected SNP had a causal effect.

	SNP	microRNA	log10 pvalue	location	demonstrated processing effect	trait implicated SNP
1	rs73933241	miR-641	11.5	flank-3p	n	n
2	rs817478	miR-4423	5.6	flank-3p	n	n
3	rs174561	miR-1908	25.3	flank-5p	n	y[238, 239]
4	rs8054514	miR-3176	18.1	flank-5p	n	n
5	rs7247222	miR-3188	10.2	flank-5p	n	n
6	rs641071	miR-4482-1	9.4	flank-5p	n	n
7	rs7911488	miR-1307	32.6	loop	n	y[240, 241]
8	rs745666	miR-3615	8.2	loop	n	n
9	rs2910164	miR-146a	36.0	miRNA-3p	y[197, 242]	y[243, 195, 242, 196]
10	rs28664200	miR-1255a	26.9	miRNA-3p	n	n
11	rs2273626	miR-4707	26.8	miRNA-3p	n	n
12	rs2155248	miR-1304	24.1	miRNA-3p	n	n
13	rs7227168	miR-4741	9.1	miRNA-3p	n	y[244]
14	rs2682818	miR-618	7.4	miRNA-3p	y[245]	y[246, 247]
15	rs2168518	miR-4513	18.4	miRNA-5p	y[248]	y[248, 249]
16	rs2427556	miR-941-3	17.7	miRNA-5p	n	n
17	rs9589207	miR-92a-1	16.1	miRNA-5p	n	y[250, 193, 251]
18	rs2241347	miR-3130-2	12.9	miRNA-5p	n	n
19	rs487571	miR-5680	12.1	miRNA-5p	n	n
20	rs10422347	miR-4745	6.2	miRNA-5p	n	y[252]

Table 6.1: Common microRNA cis-eQTLs, with expression associated to a mature microRNA from the hairpin that variants are within or adjacent to. The pvalue and segment of the microRNA hairpin the variant is within are listed. Where SNPs have been associated with microRNA expression or traits previously the references are shown.

6.3.4 microRNA cis-eQTLs are also trait associated

Several microRNA cis-eQTLs identified here have been previously identified as trait or disease associated SNPs, mainly from case-control studies. As molecular phenotypes are *a priori* good functional candidates for cellular and organism level traits because risk variants often act by

affecting aspects of gene regulation e.g. splicing QTLs which contribute as much to some traits as eQTLs[253], I have explored other genetic associations of these common mpQTVs.

Cancer associated

miR-146a is transcribed as a primary microRNA with mir-3142. rs2910164 (G>C) changing a G:U pair to a C:U mismatch in the stem loop is associated with increased expression of the 3p mature microRNA but decreased expression of the 5p microRNA (Fig 6.3a). This microRNA hand has been associated with the innate immune response[243] and identified as associated with colorectal cancer risk (OR:1.34 95% CI 1.15-1.67) and survival (Hazard Ratio 2.12)[195] and breast cancer risk (OR:1.77, 95% CI 1.40-2.24)[196]. Selection during the evolution of metastasis has also been observed in two studies in differing ways: Jazdzewski et al. [197] observed selection for the heterozygous state at this variant, with mutations from either the GG or CC homozygote in 14 of 300 tumour/normal samples (4.6%). Forloni et al[242] observed selection for the G allele for during the evolution of melanoma where in 8 of 15 cases a C allele in the primary melanoma were mutated to G alleles in the primary metastasis including heterozygote to GG mutations.

In colony forming assays melanoma cells stably expressing a miR-146a transcript with the G allele formed more colonies than those with the C allele. This effect is suggested to be due to the increased expression of mir-146a-5p which targets the NUMB mRNA - a regulator of Notch signalling. Notch signalling having a key role in melanocyte development, usually decreased in mature melanocytes and reactivation of Notch signalling required for melanoma formation[242]. This locus also appears to be under strong selection with the ancestral G allele conserved across most placental mammals however the C allele is at a high frequency in human populations. Also there appears to be fewer heterozygotes for this SNP than would be expected under Hardy-Weinberg equilibrium although for the 452 samples genotyped here the odds ratio of 0.786 does not reach statistical significance (p=0.096).

miR-618 is located within an intron of LIN7A (a cell polarity complex component). rs2682818 (C>A) located in the 3p microRNA is associated with a decrease in the expression of the 5p mature microRNA (Fig 6.3b). miR-618 has been observed as up-regulated in a number of cancers including hepatocellular carcinoma and male breast cancer[246, 247]. Genotyping of this SNP in in non-Hodgkin lymphoma (455 cases, 527 controls) found an elevated risk for individuals with one or more copies of the A variant (OR: 1.65, 95% CI: 1.05-2.60).

miR-4741 is located within the retinoblastoma binding protein 8 (RBBP8). rs7227168 (C>T) located within the 3p mature microRNA and is associated with a decrease in expression of the 3p mature microRNA (Fig 6.3m). This SNP has been associated with an increased mortality risk in non-small cell lung cancer (Hazard ratio 1.35, 95% CI 1.11-1.65)[244].

miR-1307 is located within a cassette exon of up-regulated during skeletal muscle growth 5 (USMG5). rs7911488 (A>G) located within the terminal loop is associated with a decrease

in expression of both the 3p and 5p mature microRNA (Fig 6.3q 3p shown as it is most significant). This SNP has been associated with occurrence of colorectal cancer (OR: 1.29, 95% CI 1.13-1.46)[241] and has a proposed mechanism through the C allele creating a binding site for exonic splicing enhancers MBNL1 which may either block Dicer processing[241] or affect exon skipping[240].

miR-92a is part of the miR-17∼92 microRNA cluster which has been observed to have amplifications and overexpression in cancers[250] and deletions leading to growth defects[193]. rs9589207 (G>A) located in the 5p mature microRNA is associated with an increase in expression of the 5p mature microRNA (Fig 6.3t). The SNP has been associated with decreased risk of gastric cancer with a possible mechanism in altering the binding to FBXW7 - a protein involved in cell proliferation which has been related to tumour pathogenesis - leading to upregulation in presence of the AA genotype[251].

Blood lipid traits

Examining trait associations based on expression in LCLs will miss many associations relevant to other cells and tissues, however blood lipid traits may be the most relevant traits to this cell type.

miR-1908 is located within an intron of FADS1 (fatty acid desaturase 1A). rs174561 (T>C) located in the 5p basal stem is associated with an increase in expression of the mature 5p and 3p microRNA (Fig 6.3d 5p shown as this association was most significant). mir-1908 is known to be highly expressed in mature human adipocytes with expression level affected by obesity and insulin sensitivity factors[238]. rs174561 has been associated with LDL-cholesterol, HDL-cholesterol, total cholesterol, triglyceride, and fasting glucose. However nearby SNPs in LD with rs174561 were also associated with these traits making the identification of the causal variant difficult[239].

miR-4513 is located within an intron of CSK (c-src tyrosine kinase). rs2168518 (G>A) located within the seed region of miR-4513-5p is also associated with a decrease in expression of the 5p mature microRNA (Fig 6.3f). This SNP has been associated with fasting glucose, low-density lipoprotein, cholesterol, total cholesterol, systolic and diastolic blood pressure, and risk of coronary artery disease[248, 249].

Other traits

miR-4745 is contained within an intron of he PTBP1 (polypyrimidine tract binding protein) adjacent to an exon and potentially processed as a mirtron where the pre-microRNA is excised by the spliceosome. rs10422347 (C>T) located at position 18 of the 5p microRNA is associated with a decrease in the expression of the 5p microRNA (Fig 6.3c). However other

nearby microRNAs are more strongly associated with expression. rs10422347 is in LD with SNPs in the PTBP1 gene associated with reduced insulin release[252].

6.3.5 SNPs with known microRNA processing effects are recovered

Within these identified cis-eQTLS were several cases of SNPs with experimentally validated effects on microRNA processing rather than pri-microRNA production. Although, without exception the mechanistic basis of that processing change has not yet been demonstrated.

miR-618

The effect of rs2682818 on the mature microRNA level of miR-618 described above as a risk factor in several cancers and is shown in Fig 6.3b has been observed in HeLa cells where transfection with both G and T allele microRNA precursor expression vectors had approximately similar effects on the levels of precursor transcripts (10.6- and 9-fold increase) whereas the G allele expression vector increased the mature microRNA transcript levels 5-fold the T allele vector led to an increase of mature microRNA 2-fold, suggesting an effect in one of the post-transcriptional processing steps[245].

miR-146a

The effect of rs2910164 on expression of miR-146a-5p described above as selected for in colorectal cancers and shown in Fig 6.3a has been seen previously seen in an *in vitro* processing assay where in HeLa nuclear extract pre-miR-146a was produced at a two fold higher level from a G allele primary transcript compared with a C allele primary transcript, demonstrating a differences between the two alleles in the efficiency of pre-microRNA production in the nucleus[197]. This effect has been replicated by Forloni et al. who observed a higher level of miR-146a using an expression vector with the G allele than with the C allele in three different cell types[242].

miR-4513

The effect of the mutant A rs2168518 allele on lowering the production of miR-4513 described above associated with various blood glucose traits and is shown in Fig 6.3f has also been demonstrated through transfection of expression plasmids with wild type or mutant pre-miR-4513 and GFP into HEK293 cells assaying the levels of these by qPCR[248].

6.3.6 A reciprocal response from miR-146a to a cis-eQTL

rs2910164 in miR-146a discussed above as a cancer associated associated processing variant at a locus under strong selection in the human population shown in Fig 6.3a has previously been described as increasing the expression of the mature 5-prime microRNA[197, 242]. This effect could also be seen in this study with the median G/G genotype expression 9% higher than the median C/C genotype expression, however the result was not significant after multiple testing correction. A 3p microRNA from this locus was also seen with three orders of magnitude lower expression than the 5p microRNA (miR-146a-5p being the 4th most abundant microRNA). rs2910164 was seen as a significant eQTL for miR-146a-3p after multiple testing correction, with the C/C genotype having expression about double that of the G/G genotype.

This result suggests a reciprocal response to the SNP where the G allele is associated with an increase of the 5p microRNA (as has been experimentally validated) and a decrease of the 3p microRNA, suggesting a novel possibly competitive effect in the processing of this hairpin. The effect of SNPs on microRNA expression where there is expression data for both 5p and 3p products is shown in Fig 6.4, for cis SNPs within 1kb of a microRNA. The correlation of rs2910164 and another SNP in strong LD with miR-146a was unique in showing an opposite response by the 5p and 3p microRNAs. All other cis-eQTL SNPs (FDR$<$1%) the correlation with expression was always in the same direction for both the 5p and 3p microRNA.

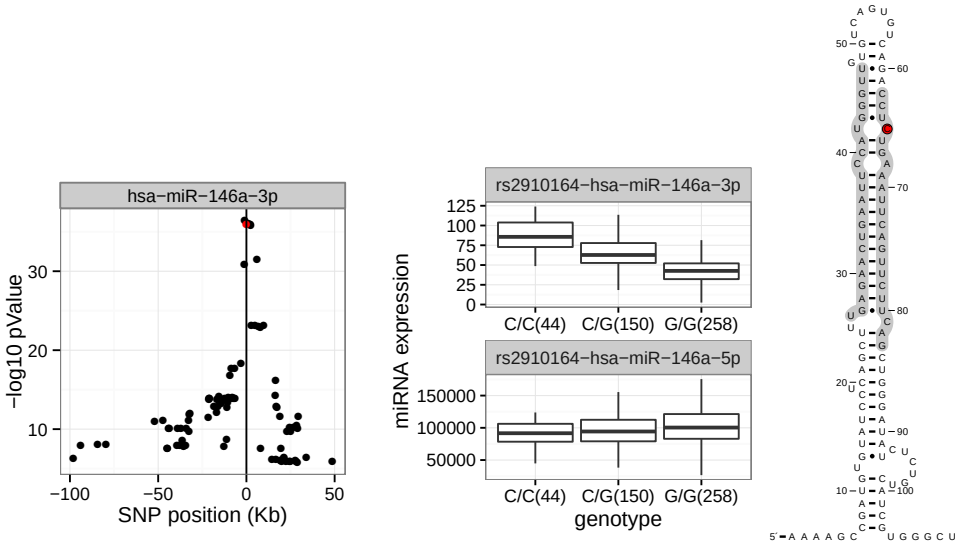
6.3.7 Rare variants are cis-eQTLs

Common variants are unlikely to manifest as highly penetrant deleterious phenotypes as selection will tend to reduce their frequency within the population. Consequently, rare SNP eQTLs may be more likely to have large effects expression levels and subsequently more deleterious effects on phenotypes, although the power to detect them is lower than for common variants.

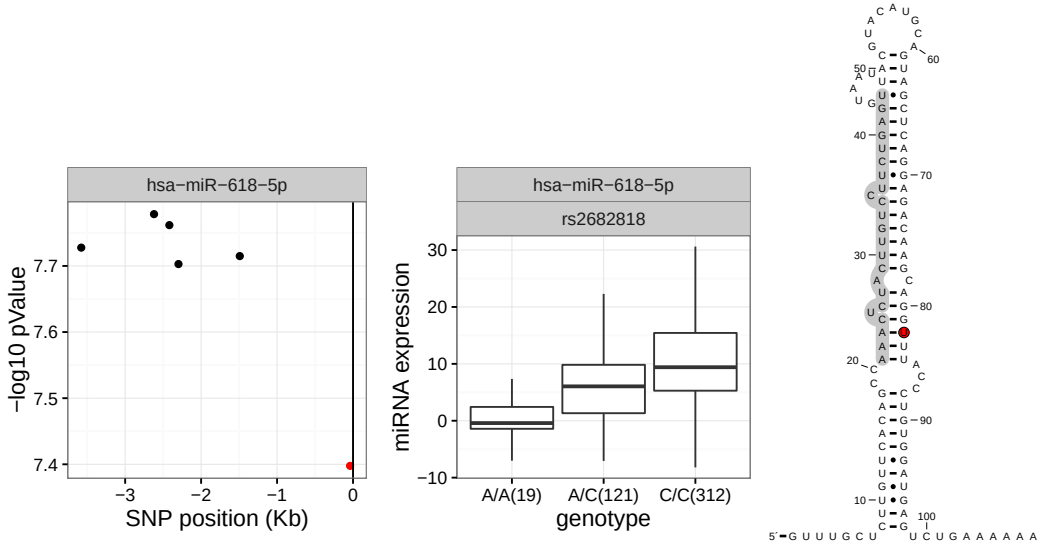
This analysis has identified 12 rare microRNA cis-eQTLs which were present in either the microRNA terminal loop or flanking the mature microRNA (Table 6.2.

Boxplots showing the effect on expression for these 12 cis-eQTLs are shown in Fig 6.5, and locus plots are shown in Fig 6.6 for 11 regions as two SNPs are selected as cis-eQTLs for miR-769-5p. In some cases such as the SNPs selected for miR-106b and miR-576 the selected SNP in the hairpin loop is the most significant within the region. However in other cases other similarly significant or more significant cis-eQTL SNPs are present near the microRNA locus. In these later cases performing multivariate linear regression using the genotype values of multiple SNPs to explain the expression changes allows us to see if one SNP has a significant effect in explaining expression when the effects of other SNPs have been taken into account.

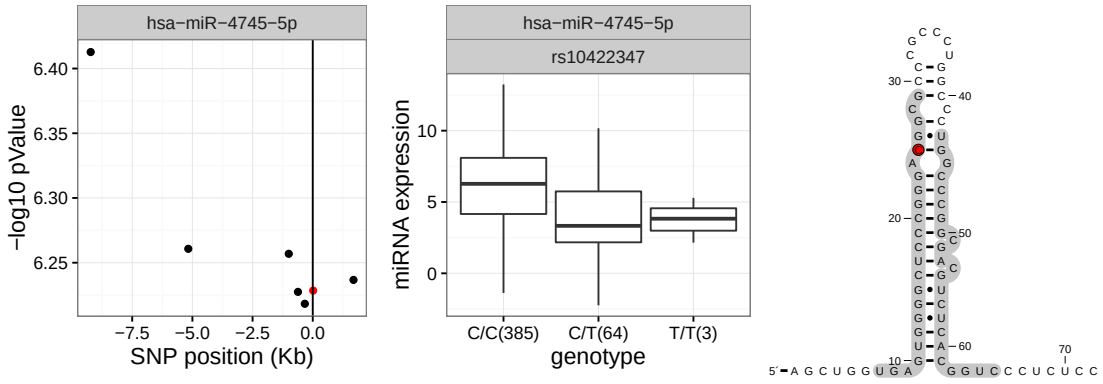
For example in the case of miR-5701 a SNP 150bp downstream of the locus is most significant with a -log10 pvalue of 3.7 whereas the SNP within the microRNA hairpin has a -log10 pvalue



(a)

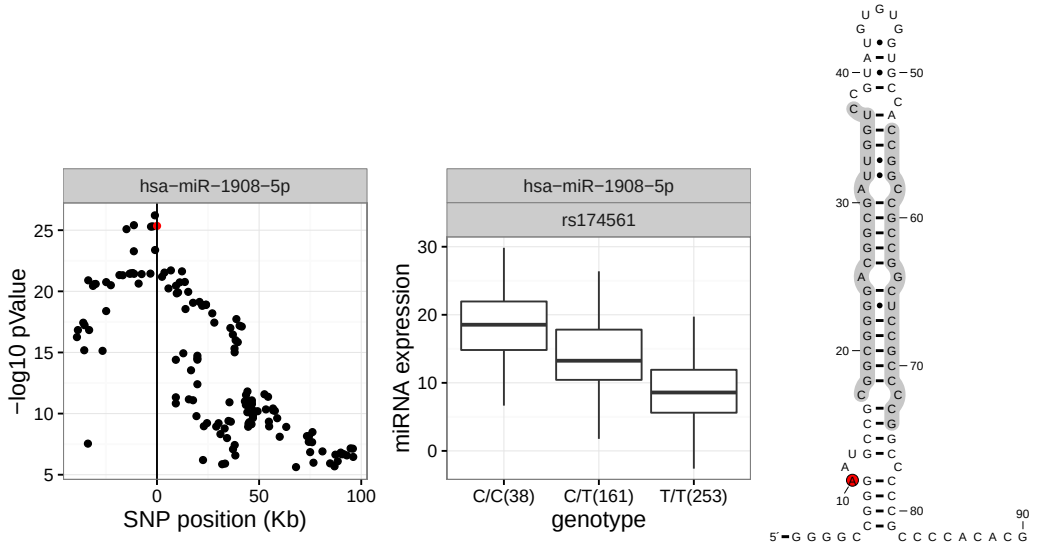


(b)

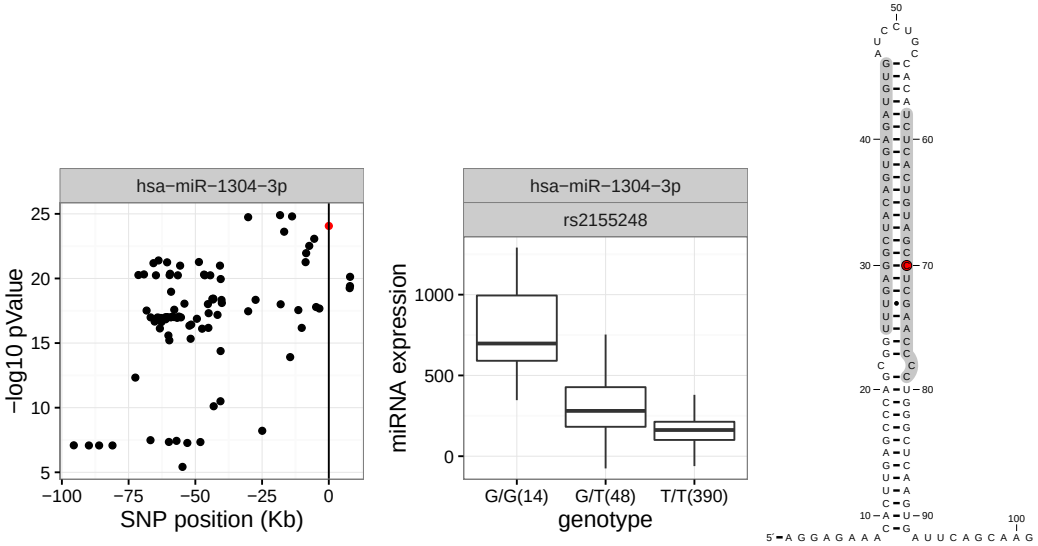


(c)

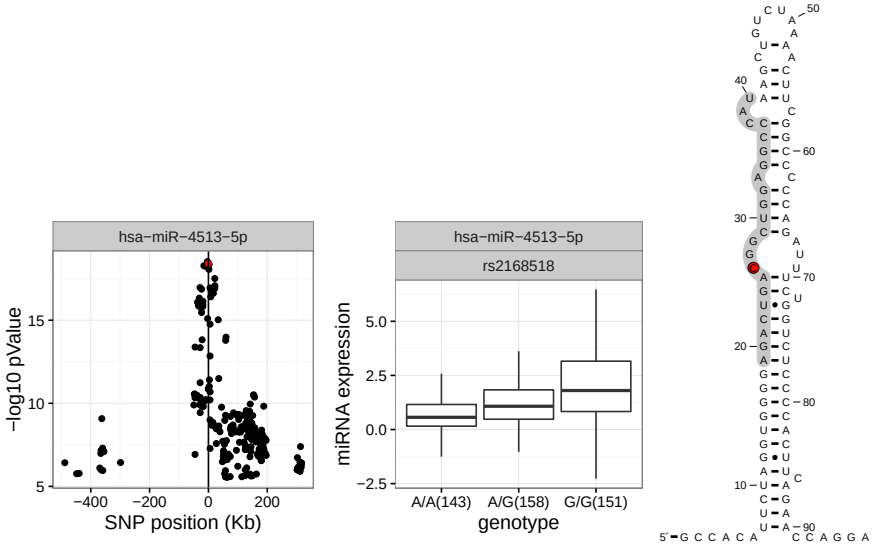
Figure 6.3: Left: Regional association plots (x-axis) genomic distance from microRNA, (y-axis) -log10 pvalue. Centre: microRNA expression boxplots (x-axis) genotypes with number of samples, (y-axis) normalised microRNA expression level. Right: schematic of the microRNA locus with mature microRNA(s) shown in grey, variant shown in red.



(d)

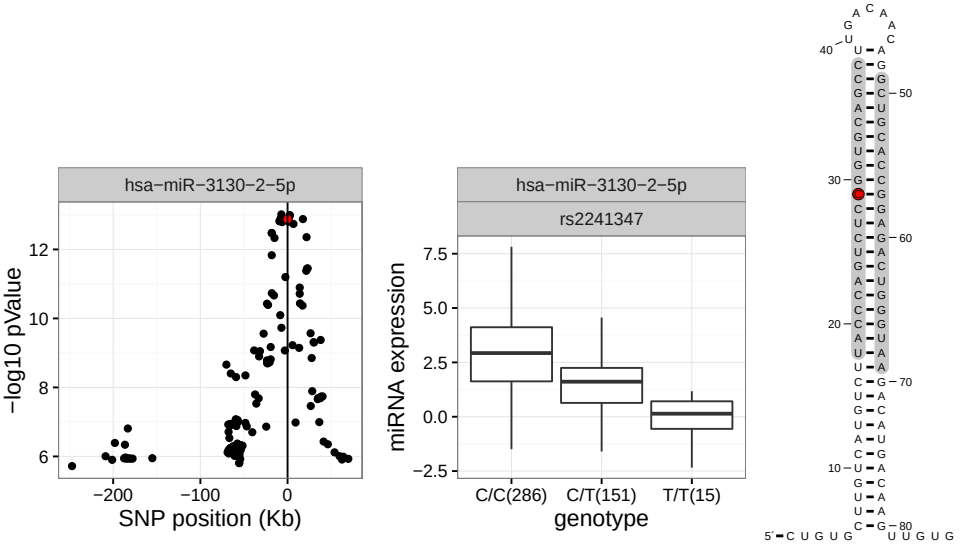


(e)

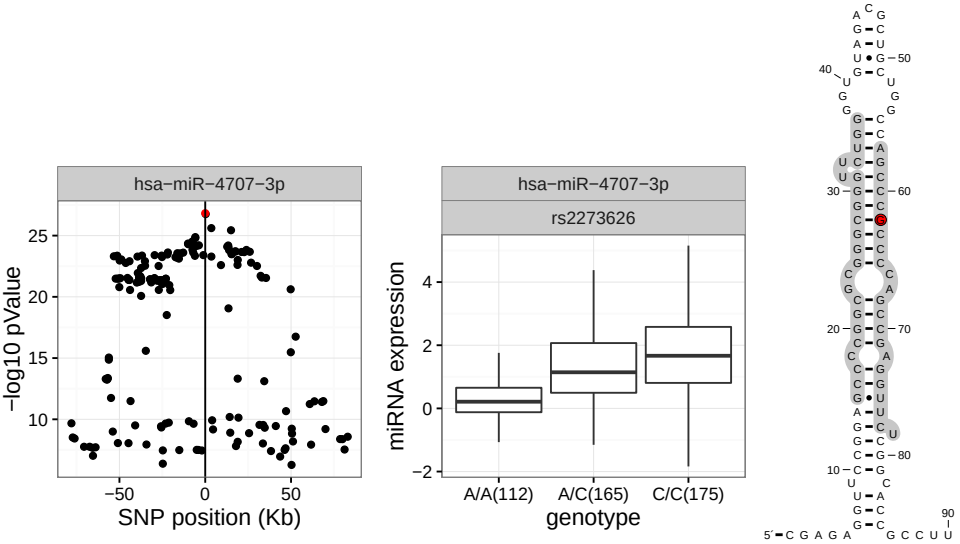


(f)

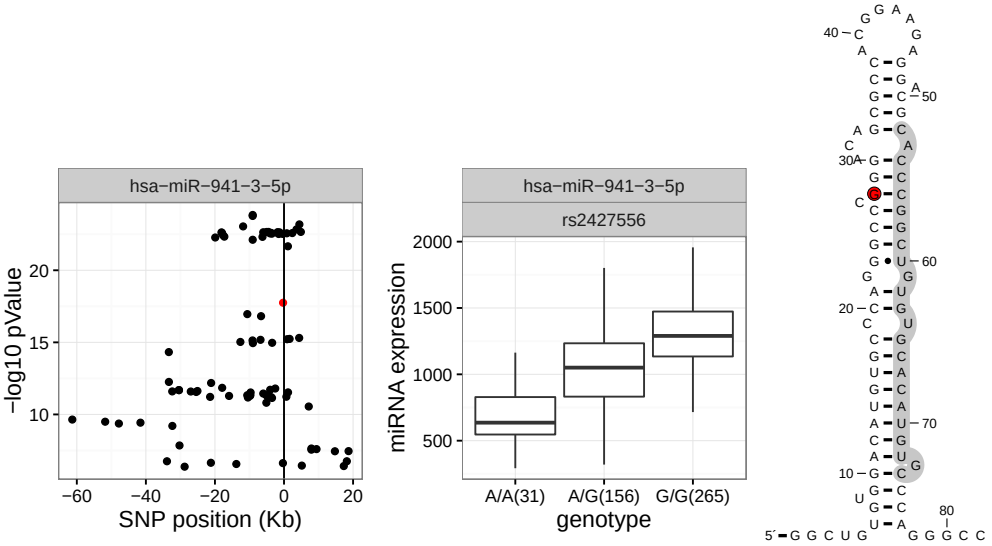
Left: Regional association plots (x-axis) genomic distance from microRNA, (y-axis) -log10 pvalue. Centre: microRNA expression boxplots (x-axis) genotypes with number of samples, (y-axis) normalised microRNA expression level. Right: schematic of the microRNA locus with mature microRNA(s) shown in grey, variant shown in red.



(g)

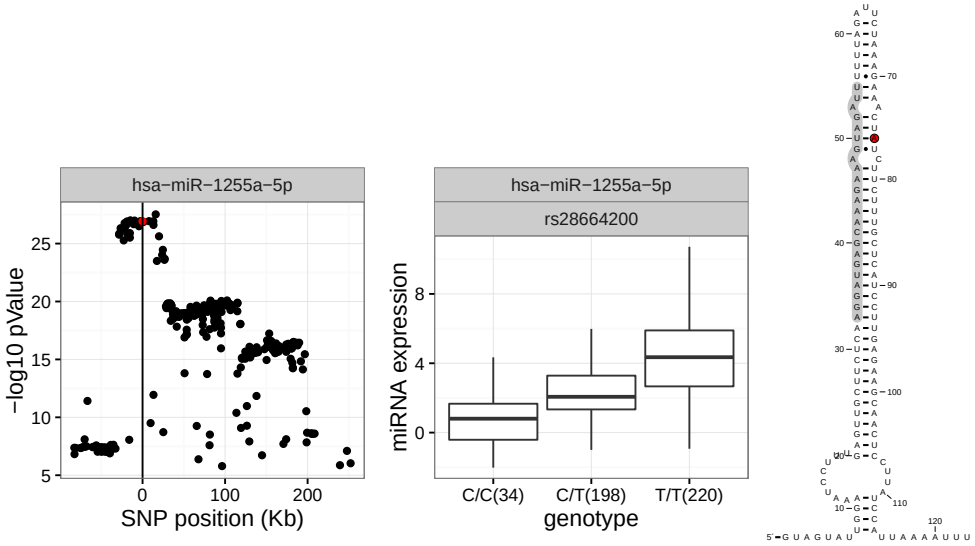


(h)

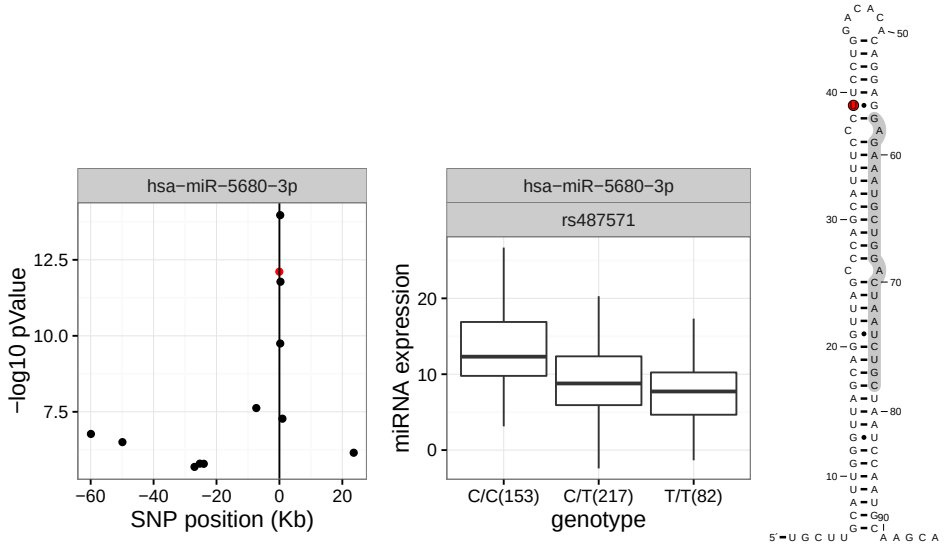


(i)

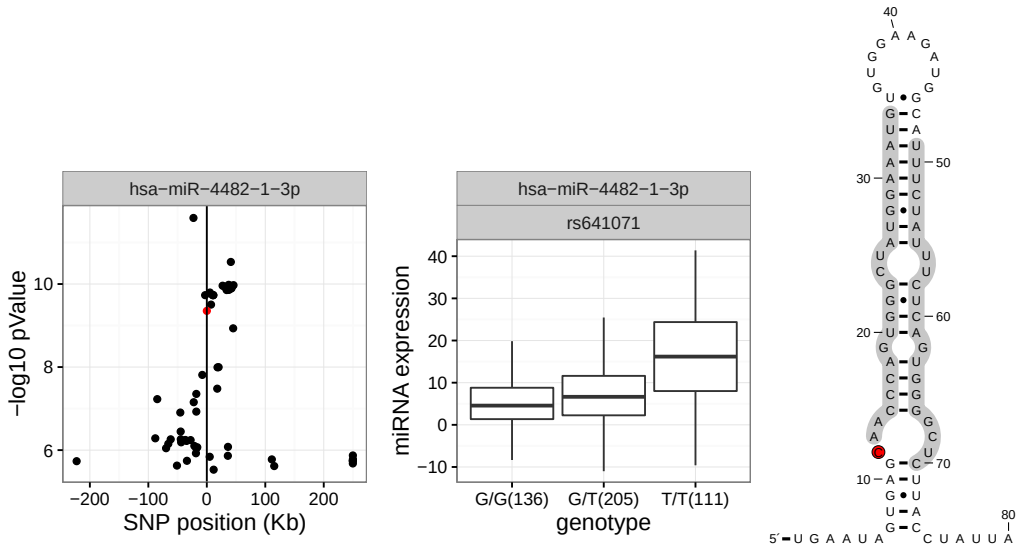
Left: Regional association plots (x-axis) genomic distance from microRNA, (y-axis) -log10 pvalue. Centre: microRNA expression boxplots (x-axis) genotypes with number of samples, (y-axis) normalised microRNA expression level. Right: schematic of the microRNA locus with mature microRNA(s) shown in grey, variant shown in red.



(j)

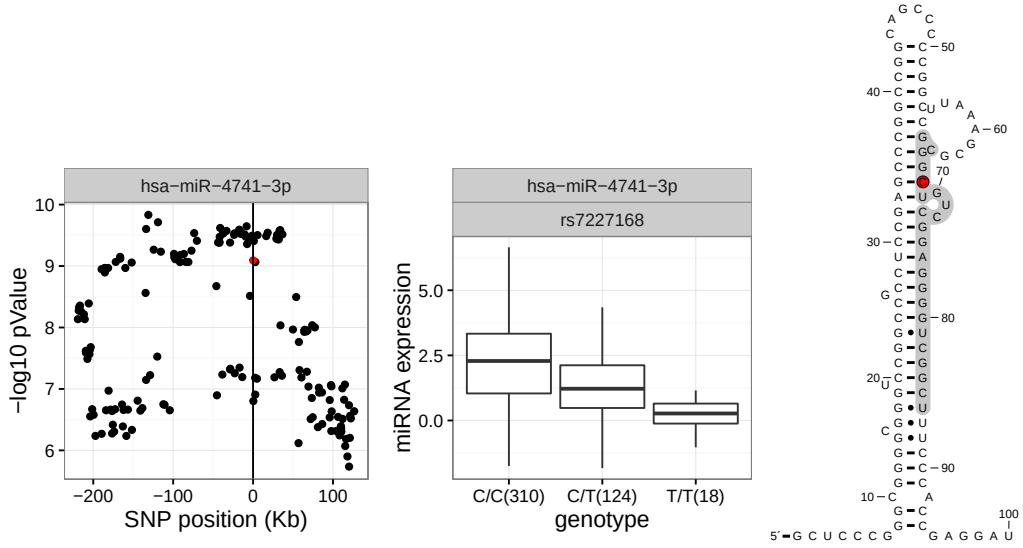


(k)

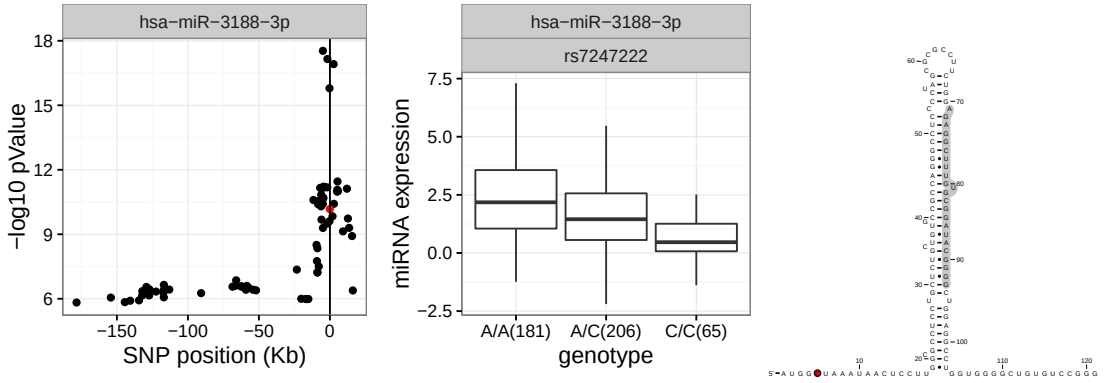


(l)

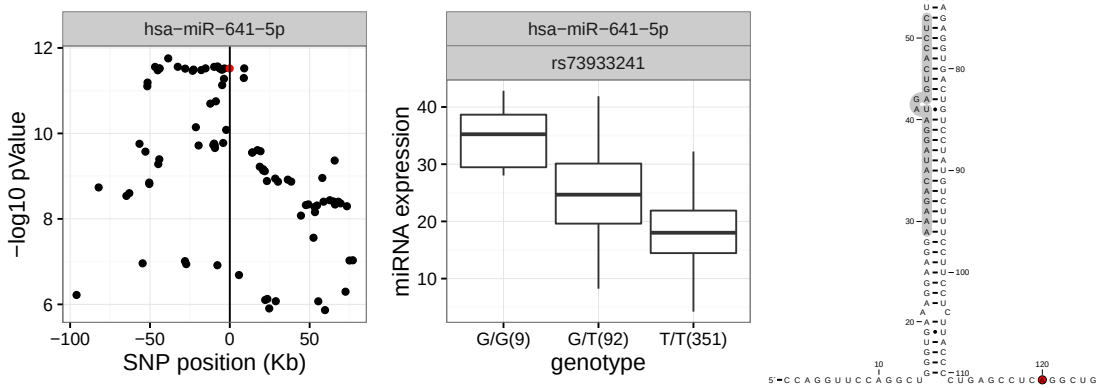
Left: Regional association plots (x-axis) genomic distance from microRNA, (y-axis) -log10 pvalue. Centre: microRNA expression boxplots (x-axis) genotypes with number of samples, (y-axis) normalised microRNA expression level. Right: schematic of the microRNA locus with mature microRNA(s) shown in grey, variant shown in red.



(m)

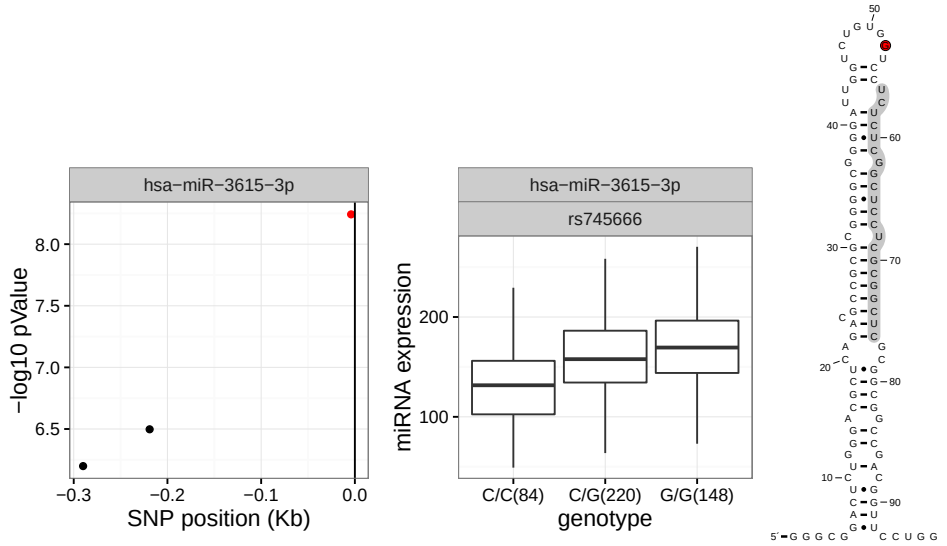


(n)

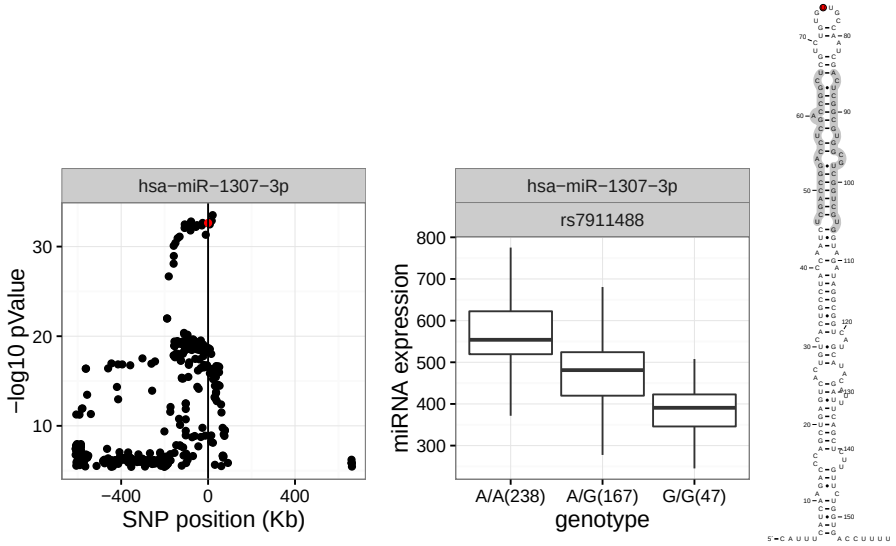


(o)

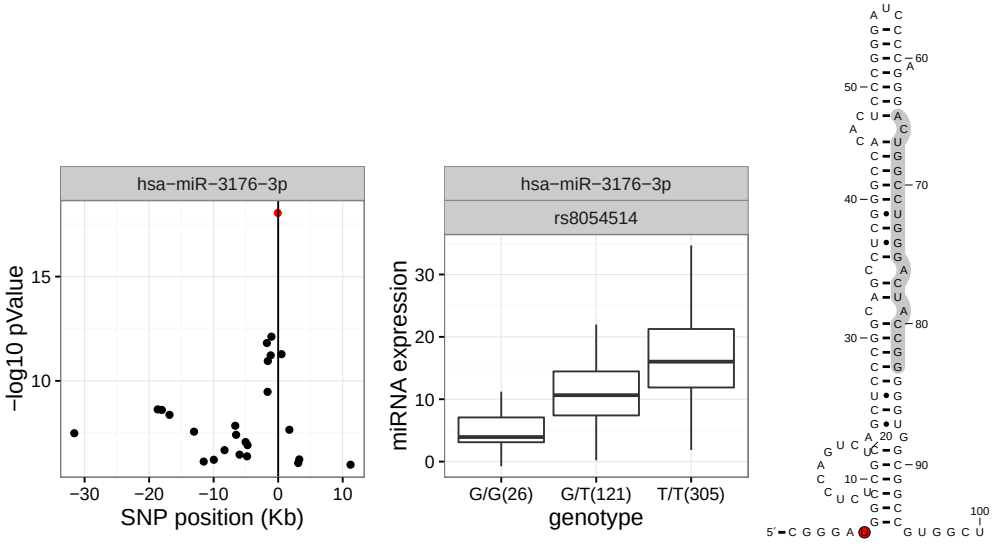
Left: Regional association plots (x-axis) genomic distance from microRNA, (y-axis) -log10 pvalue. Centre: microRNA expression boxplots (x-axis) genotypes with number of samples, (y-axis) normalised microRNA expression level. Right: schematic of the microRNA locus with mature microRNA(s) shown in grey, variant shown in red.



(p)

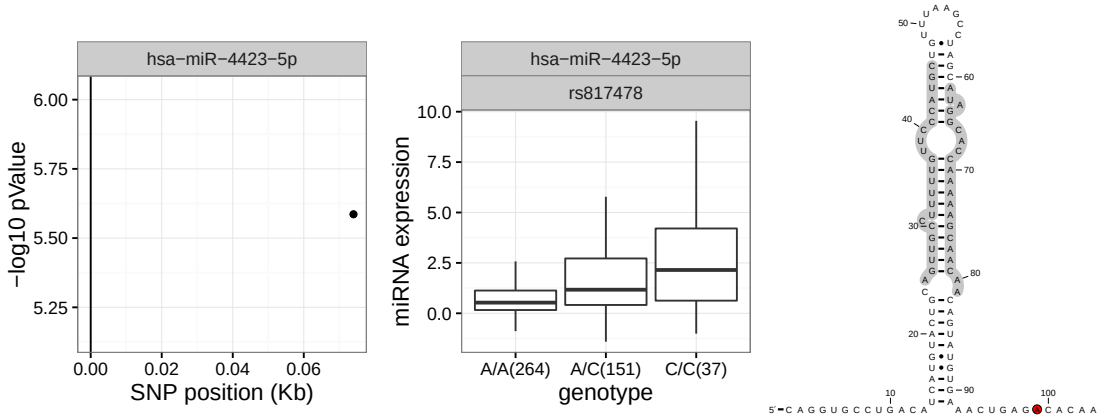


(q)

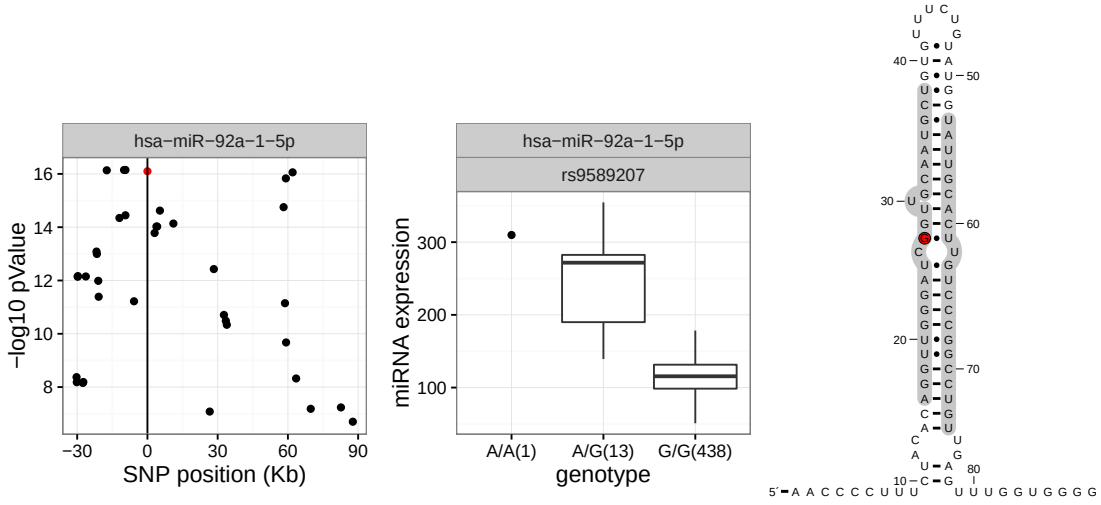


(r)

Left: Regional association plots (x-axis) genomic distance from microRNA, (y-axis) -log10 pvalue. Centre: microRNA expression boxplots (x-axis) genotypes with number of samples, (y-axis) normalised microRNA expression level. Right: schematic of the microRNA locus with mature microRNA(s) shown in grey, variant shown in red.



(s)



(t)

Left: Regional association plots (x-axis) genomic distance from microRNA, (y-axis) -log10 pvalue. Centre: microRNA expression boxplots (x-axis) genotypes with number of samples, (y-axis) normalised microRNA expression level. Right: schematic of the microRNA locus with mature microRNA(s) shown in grey, variant shown in red.

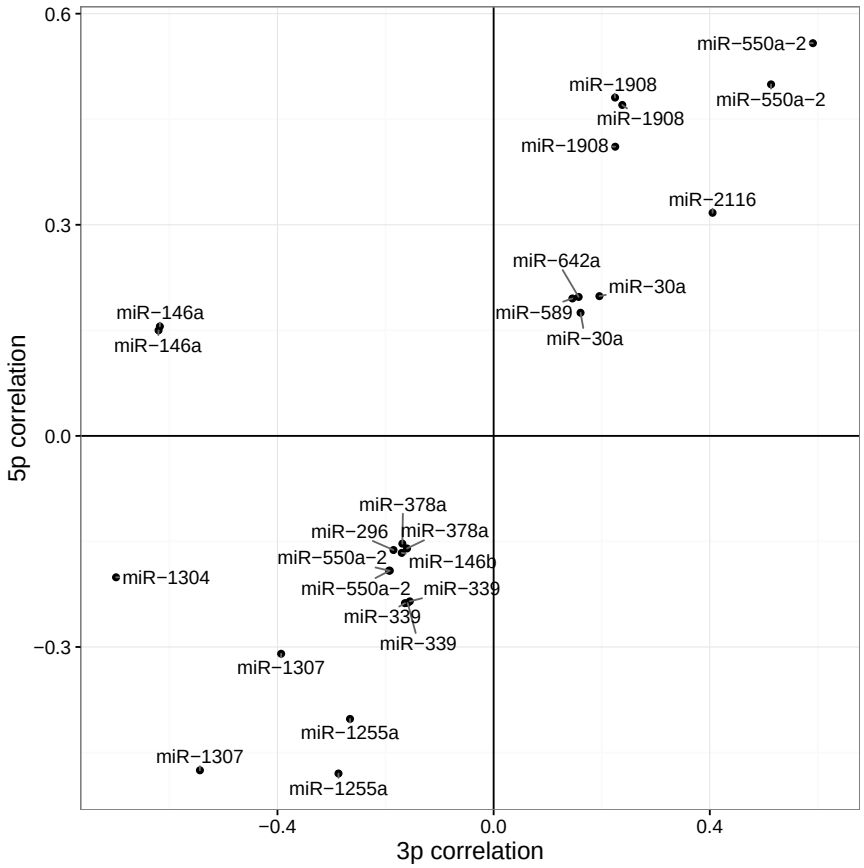


Figure 6.4: Correlation of variants with 3p and 5p mature microRNA expression from the same locus within 1kb. (X-axis) correlation with the 3p mature microRNA. (Y-axis) correlation with the 5p mature microRNA.

	microRNA	chr	pos	maf	-log10(pvalue)	location
1	miR-4786-5p	2	240882519	0.003	1.676	flank-5p
2	miR-576-5p	4	110409962	0.008	2.896	flank-3p
3	miR-2277-5p	5	92956409	0.029	5.325	stem_3p
4	miR-1303-3p	5	154065348	0.030	2.886	stem-5p
5	miR-1303-5p	5	154065348	0.030	2.436	stem-5p
6	miR-589-5p	7	5535434	0.011	4.528	flank-3p
7	miR-106b-3p	7	99691652	0.008	2.890	loop
8	miR-96-5p	7	129414514	0.004	2.396	flank-3p
9	miR-5701-1-5p	15	21145671	0.068	2.200	flank-3p
10	miR-4513-5p	15	75081115	0.009	3.021	flank-5p
11	miR-769-5p	19	46522189	0.004	2.488	flank-5p
12	miR-769-5p	19	46522298	0.004	2.488	stem_3p

Table 6.2: Rare microRNA cis-eQTLs identified in the GEUVADIS dataset using linear regression with the R package MatrixeQTL. Minor allele frequency(maf), pvalue and location of the variant in the microRNA hairpin shown.

of 2.2. Multivariate regression demonstrates that the effect of the hairpin SNP disappears when the most significant is taken into account, as these SNPs are in linkage disequilibrium (r^2 = 0.6).

In the case of miR-96-5p the most significant SNP is located 750bp downstream of the microRNA locus with a -log10 pvalue of 3.3 whereas the SNP in the hairpin has a -log10 pvalue of 2.4 (Fig 6.6). The regression analysis demonstrates that these SNPs are independent and the significance of the hairpin SNP remains after the more significant SNP is taken into account.

These selected rare SNPs increase the number of potential mpQTV and increase the power for subsequent analysis of precursor regions and motifs to determine which are most often altered in mpQTVs.

6.3.8 Primary hairpins are enriched for cis-eQTLs

SNPs within the whole stem-loop structure have an odds ratio of 5.6 for being a cis-eQTL compared to SNPs flanking hairpins up to 1kb this confirms my prior expectation that loci within the hairpin are more likely to affect microRNA levels than those without.

Examining each segment of the hairpin individually it can be seen in Fig 6.7 that several segments within the stem seem to contribute to this signal, despite the small numbers of SNPs in some segments in this analysis, leading to large error bars. SNPs in the 5p stem, hairpin loop, and mature microRNAs having odds ratios of 14-28 (20-45% of SNPs in these regions are cis-eQTLs) with error bars which do not overlap the null expectation demonstrating that these regions are enriched in microRNA cis-eQTLs.

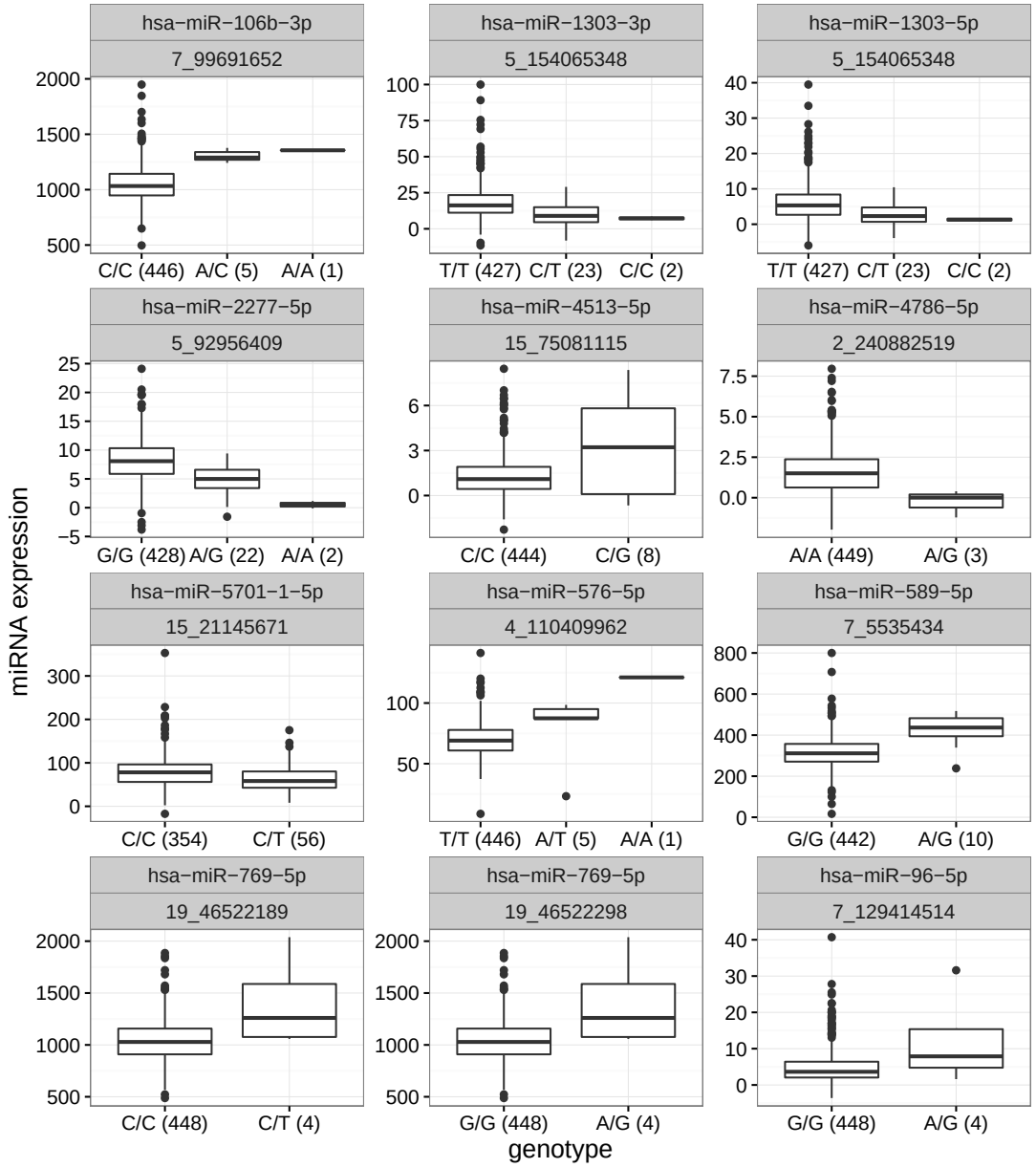


Figure 6.5: Rare variant cis-eQTL associations. (x-axis) Genotype and number of individuals. (y-axis) microRNA expression (normalised read count). Variants are labelled by chromosome_position

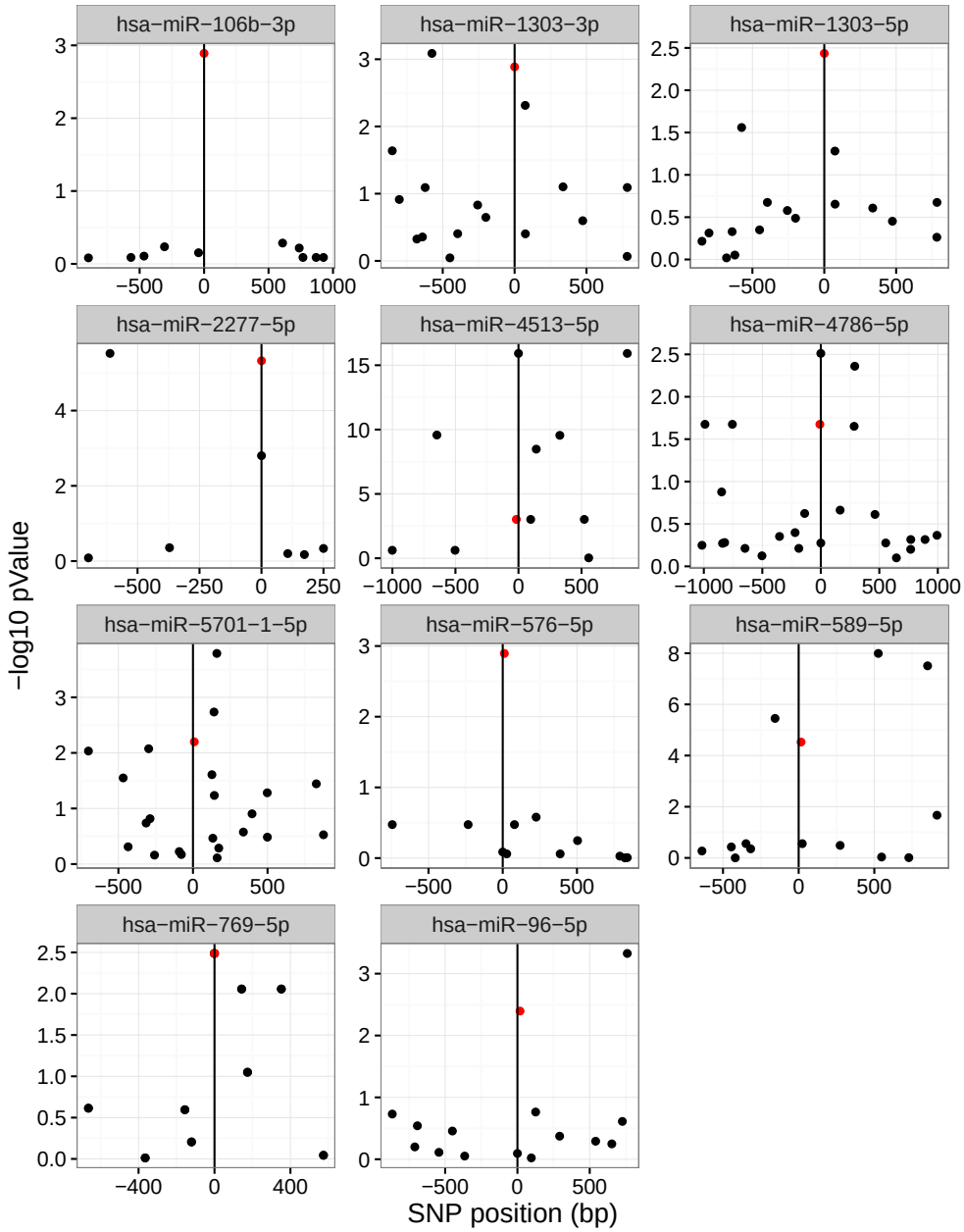


Figure 6.6: cis-eQTLs within the region for each selected microRNA. (x-axis) distance from microRNA (bp). (y-axis) -log10 P value for each SNP. Red points: Selected rare cis-eQTL SNPs within primary microRNAs. Black: Other SNPs

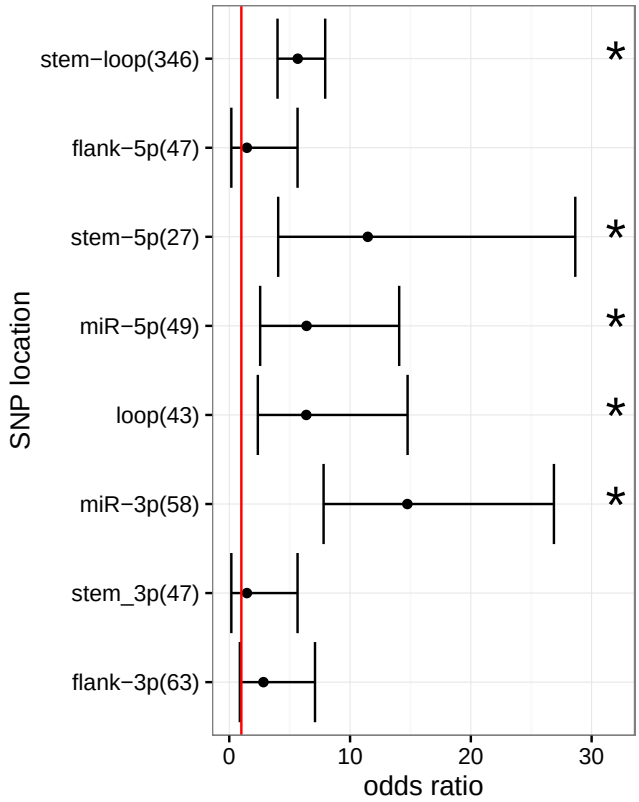


Figure 6.7: Odds ratio of containing microRNA cis-eQTLs for different regions of microRNA hairpins. Calculated by Fisher's exact test comparing to SNPs within 1kb of microRNA hairpin loci excluding those within. Number of SNPs is shown in parentheses. Asterisks denote significance $p<0.05$ with Bonferroni multiple testing correction for n=8 tests.

6.4 Chapter summary

SNPs are present across microRNA loci, and derived allele frequency tests demonstrate evidence for purifying selection above the genome-wide background at microRNA loci. Previous studies have identified a number of variants associated with microRNA expression levels but few have been investigated to identify causality and determine the mechanisms involved. I have identified candidate microRNA processing variants from common and rare polymorphisms which are cis-eQTLs for the microRNA loci they are within or adjacent to using publicly available whole-genome and short RNA sequencing data in LCLs from the 1000 Genomes Project and the GEUVADIS consortium.

Despite the fact that the discovery of potential mpQTVs in this analysis is limited to those microRNAs commonly expressed in LCLs a number of SNPs previously identified as expression QTL or trait associated SNPs were seen. This includes rs2910164 in miR-146a which has been observed to affect expression of the mature microRNA and is recapitulated as a frequent somatic mutation in several cancer types, and rs2682818 in miR-618 which has also been previously identified as a microRNA eQTL and risk factor in several cancers. rs2910164 in miR-146a was also seen to have a reciprocal relationship to the 3p and 5p microRNA products suggesting a novel competitive mechanism.

These candidate processing variants will require experimental validation to determine whether the selected SNP is indeed causal for the expression change. Where this is the case the mechanisms of these effects may be investigated. The mechanism in the case of miR-146a with 3p and 5p mature microRNAs responding in opposite directions to the presence of the eQTL SNP may be particularly novel.

Stratifying microRNA cis-eQTLs by the segment of the microRNA hairpin they are contained within suggests that SNPs within the 5p-stem, mature microRNAs and terminal loop of the hairpin are more likely to affect the expression level of the mature microRNA, leading to phenotypic consequences.

6.4.1 Future Directions

These candidate microRNA processing variant SNPs are being investigated in microRNA processing assays by our collaborators to determine if they are the causal variant, and if so which stage of processing they affect: the cleavage in the nucleus by DROSHA/DGCR8, export, or cleavage in the cytoplasm by DICER. Following on from this protein binding assays and RNA structure analysis such as SHAPE and hydroxyl radical cleavage footprinting as used by Fernandez et al[199] (section 1.7.4) can determine where effects are due to the disruption or creation of protein binding motifs or changes in RNA structure. This analysis will lead to a greater understanding of microRNA processing mechanisms and the effect that human polymorphisms have on this process.

Where short RNA-seq and genome sequencing data are available in other - perhaps disease specific - cells or tissues similar analyses could be performed to identify additional mpQTVs, particularly for microRNAs not expressed in LCLs and in disease or trait relevant cell types.

As microRNAs are involved in the regulation of many cellular and developmental pathways, and variants at microRNA loci may alter microRNA expression, common polymorphisms and rare mutations at microRNA loci have been seen to predispose to certain cancers and to cause mendelian traits (section 1.7.3). These loci are often ignored, not prioritised in genome sequencing studies or variant prioritisation software. Chapter 7 will examine microRNA variants in the context of exome sequencing studies of disease cohorts.

Chapter 7

MicroRNA variants in disease cohorts

7.1 Introduction

microRNA disruption and disregulation has been implicated in a variety of diseases through; expression levels being associated with disease outcomes, common polymorphisms being associated with disease outcomes in case-control studies, and rare variants in two causing developmental phenotypes (section 1.7.3).

microRNA loci are non-protein-coding regions of the genome which are frequently ignored in sequencing studies where variant annotation pipelines such as SIFT[254] and PolyPhen[255] score variants by the severity of their effect on proteins i.e. missense, nonsense, frameshift as well as annotating splice site and UTR mutations. Predictive information about the consequences of microRNA variation is lacking and few examples of microRNAs disrupted in disease have been observed. However microRNA processing defects have a clear molecular model whereby a change in microRNA expression can impact the expression of other identifiable genes. These effects on microRNA production can be tested experimentally in processing assays, and effects on target gene expression can be tested through transfection with microRNA mimics or inhibitors followed by assaying gene expression.

Here I annotate microRNA variants in disease specific exome sequencing cohorts at the IGMM with clinical collaborators willing to perform follow up analyses and access to patient samples including larger collections of phenotypically similar patients. I identify microRNAs enriched for rare variants in particular diseases, rare variants in microRNA loci which are predicted to target genes which when disrupted are disease causing, potential pathogenic variants, and examine the profile of variants around microRNA hairpins.

7.2 Methods

7.2.1 Data

An in-house database of exome sequencing variants from disease cohorts studied at the IGMM was used. Samples are exome sequencing of normal tissue in colorectal cancer, microcephalic osteodysplastic primordial dwarfism (MOPD), cranio-facial malformation types including Cornelia de Lange syndrome collectively called 'eye', micro syndrome and myopia cohorts. These variants are stored in a MySQL database which may be queried to select genes, variants, or regions of interest. Numbers of samples for each disease cohort are shown in table 7.1. I query this database selecting variants +/- 20bp of microRNA loci from miRBase v21[11]. Also as was performed by the GEUVADIS consortium[186] microRNA hairpins where only one mature microRNA was annotated a partner microRNA was defined through base pairing after RNA structure prediction. This allows all microRNA loci to be annotated similarly, making comparisons between loci consistent.

For the MOPD cohort some sequencing of trios (parents plus affected offspring) was performed allowing me to screen for *de novo* mutations which are present in the affected offspring but not the parents.

Project	Samples
Eye developmental disorders	241
Microcephalic osteodysplastic primordial dwarfism (MOPD)	332
Micro syndrome	35
Myopia	15
Coleorectal Cancer	672

Table 7.1: Exome samples examined for mutations at microRNA loci

Target capture for exome sequencing in these samples had been performed using several different platforms (TruSeq, AgilentV4, AgilentV5, Agilent38M, Agilent50M, SeqCapv3) each with differing genomic coverage. For this reason minor allele frequency calculations and comparisons to control populations will require taking into account the number of samples from each cohort with coverage for each specific microRNA. For each microRNA I have calculated the expected coverage in each disease cohort to give microRNA specific enrichments and statistics.

Within the disease cohorts examined here several different modes of inheritance are expected: Within the rare Mendelian disorders of MOPD and cranio-facial developmental disorders causal variants are expected to be fully penetrant, either dominant *de novo* mutations or recessive rare variants forming homozygotes or compound heterozygotes. Colorectal cancer is a complex trait with likely contributions from many common and rare variants with low penetrance.

7.2.2 Filtering

Variants from the database were selected for presence within 20nt of a microRNA hairpin (miRBase v21[11]). Common variants (MAF>0.01) in the 1000 genomes project[204] or Exome Aggregation Consortium[209] (ExAC) were excluded as for the disease cohorts used common variant analysis has already been performed, and other than the cancer cohort the expected mode of inheritance is with recessive rare variants, and dominant *de novo* mutants. Variants were also excluded if coverage in a particular sample was below 10 reads or had a genotype quality score lower than 90 from the GATK variant caller[256]. Some common variants remained in the results despite this filtering, many may be population specific variants where cohorts had been acquired in a particular geographic region. For this region variants were also filtered for within-cohort minor allele frequency (MAF<0.01).

7.2.3 Variant annotation

SNPs were annotated with their microRNA locus finding overlaps for their genomic coordinates with `bedtools intersect`. Based on miRBase v21[11] annotation of microRNA processing cleavage sites and stem-loop base pairing structure, each microRNA hairpin was split into consecutive segments: 3p and 5p flanking the drosha cleavage site, 3p and 5p mature microRNA, 3p and 5p mature microRNA seed sequence, and terminal loop. SNPs were then annotated with their microRNA region accounting for microRNA strand orientation.

microRNAs were also annotated with their predicted targets from TargetScan v7[96], used to analyse microRNAs which are predicted to target disease associated genes.

7.2.4 Burden analysis

Burden analyses for enrichment of variants at individual microRNA loci and at groups of loci between cohorts are performed with Fisher's exact tests using the R function `fisher.test` from the stats package.

7.3 Results

7.3.1 microRNA loci are captured in exome sequencing

The majority of disease specific sequencing data currently available is exome sequencing. In exome sequencing a DNA capture kit with specific hybridisation targets is used to select the DNA molecules which are subsequently sequenced. This is a trade off between cost and genomic coverage compared with whole genome sequencing. Exome capture kits frequently also capture DNA from regulatory regions and non protein-coding genes, however this

coverage can be inconsistent as some regions are captured more efficiently than others. Edge-capture effects are seen where regions flanking targeted regions can be captured with low or variable coverage. Off target capture may also occur where probes hybridise to regions other than those they were designed to capture. Exome sequencing can efficiently identify SNPs and small indels in the targeted regions, however detecting larger segmental copy changes is difficult due to the non-uniformity of coverage.

microRNA coverage amongst the six different capture kits used to generate data in the in-house database is shown (Fig 7.1 and Fig 7.2). In Fig 7.1 a Venn diagram for the number of microRNA loci from miRBase v21 covered by five of the six platforms is shown. Of the 1870 microRNA loci the platforms cover 27% to 79%.

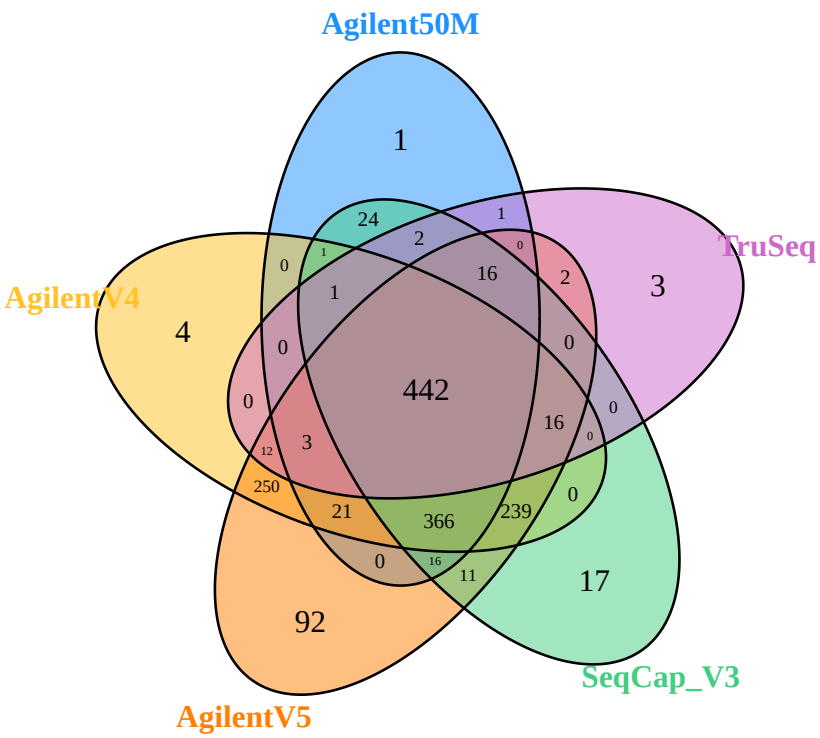


Figure 7.1: Venn diagram showing microRNA loci captured by five exome sequencing capture platforms

Amongst miRBase v21[11] annotated high confidence microRNAs - annotated due to having a high read coverage in shortRNA sequencing studies - the coverage by the target capture platforms is better, with 74% to 97% coverage of the 295 high confidence microRNAs. A Venn diagram showing the overlap in coverage for five of the six target capture platforms is shown in Fig 7.2.

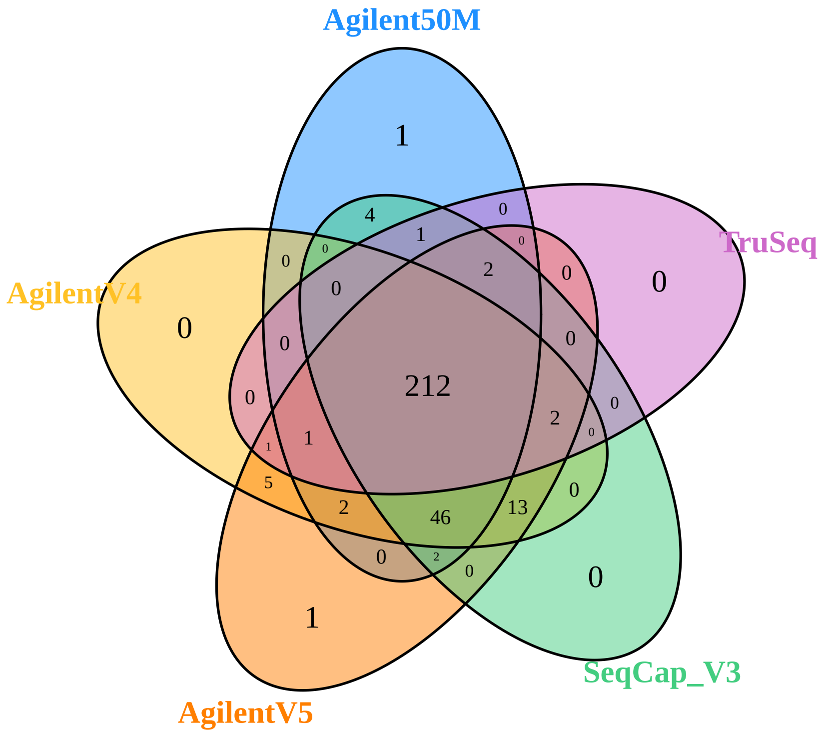


Figure 7.2: Venn diagram showing high confidence microRNA loci captured by five exome sequencing capture platforms

Filter	CRC	Eye	MOPD	micro	myopia
None	380	872	1814	424	85
Quality score / read depth	349	789	1629	371	70
ExAC & 1000g maf < 0.01	203	330	762	70	7
Cohort maf < 0.01	271	320	1025	58	6
All filters	178	191	546	28	4

Table 7.2: Variants at microRNA loci remaining after filters are applied within each cohort.

Rare and Unique variants in exome sequencing at microRNA loci

Table 7.2 shows the number of variants remaining after applying minor allele frequency and unique variant filters within each cohort. Most samples contain rare variants excluding those with MAF$>$1% within the dataset ($>$75% of samples for CRC, eye, MOPD and micro syndrome). Approximately half of samples contain variants which are unique within the pooled dataset.

cause	samples	In trio
known	73	27
unknown	102	36

Table 7.3: MOPD sample status. Causal variants had been previously identified for some of the samples. Some samples also had parental exome data.

7.3.2 Variants are distributed across microRNA loci

1660 polymorphisms at microRNA loci were present in the database after excluding known common variants, each found in between 1 and 217 individuals (median: 1, mean: 3.9). The distribution of these polymorphisms across a standardised microRNA locus is shown in Fig 7.3. Variants can be seen spread at similar frequencies across the whole length of microRNAs - flanking regions, microRNAs and terminal loops - as was also seen looking at common variants in dbSNP v139[208] (section 6.1).

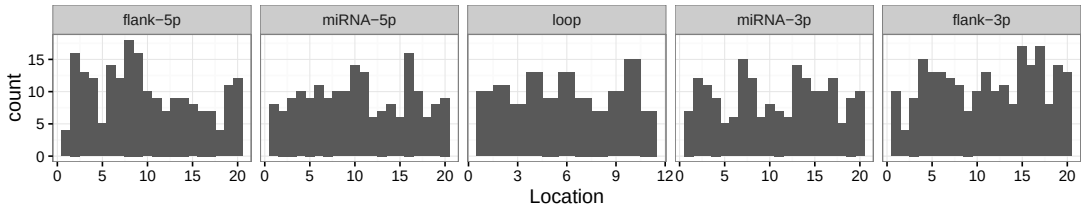


Figure 7.3: Bar chart showing the count of variants present at each position in a standardised microRNA locus

Within the rare Mendelian disorder samples it is expected that cases will generally arise due to spontaneous *de novo* mutations or rare variants segregating in the population with a recessive effect coming together as a homozygous or compound heterozygotes genotype.

In these analyses I identify *de novo* mutations where possible in the MOPD cohort, analyse homozygous and potential compound heterozygous rare variants. I also aggregate rare variants by microRNA locus and by microRNA predicted targets using Fisher's exact tests to identify microRNAs or groups of microRNAs enriched for rare variants.

7.3.3 *De novo* mutations in MOPD trios

MOPD samples were split into parental samples and cases. Cases were further split into those where the cause of the disease is unknown and those where a putative causal mutation had been identified in a known MOPD disease gene.

De novo variants could be identified through filtering for unique variants in those MOPD samples which were present in a trio, assuming equal target capture and sequencing depth in cases and parents. However as only variants added to the database are present, coverage in loci where variants are not recorded is unknown, although if a variant is present with low coverage

microRNA	region	group	cause	chr	pos	ref	alt	geno
mir-1299	flank-3p	MOPD	known	9	69002230	G	C	0/1
mir-1299	loop	MOPD	unknown	9	69002276	C	T	0/1
mir-4477a	loop	MOPD	unknown	9	68415343	A	G	0/1
mir-615	flank-3p	MOPD	unknown	12	54427841	C	CGGA	0/1

Table 7.4: MOPD putative *de novo* mutants identified for those samples with parental information where variants were present in the case sample but neither of the parental samples.

in a region it may be assumed that the region is not captured efficiently in other samples.

Four putative *de novo* mutations were seen in different MOPD cases(Table 7.4), three in cases of unknown cause and one in a case with an identified causal mutation.

chr9-69002276:C>T in the terminal loop of miR-1299 was seen in a case of unknown cause. However at the same microRNA loci chr9:69002230:G>C in the 3p-flank was also seen as a *de novo* mutation in an MOPD patient with a known causal mutation, suggesting that this variant is unlikely to be causal. Nearby rs79965448 in the 3p mature microRNA is marked in ExAC r0.3[209] as a low quality variant captured in only 257 individuals, suggesting poor coverage in a parental sample may cause its *de novo* classification here.

chr9-68415343:G>A(rs71224722) in the terminal loop of miR-4477a was present in one case of unknown cause, also seen in 47/1867 individuals in ExAC r0.3[209] where it is marked as a low quality site making it possible that this region was not well captured in the parent samples. This variant is also present in the J. Craig Venter genome[257] so not causal for a microcephaly phenotype.

A 3bp insertion at chr12:54427841 in the 3p flank of miR-615 was seen in one MOPD case of unknown cause. Other variants at miR-615 within ExAC r0.3[209] are reported as having low coverage with mean coverage in the region of 5x.

These four putative *de novo* mutations in three microRNA loci are most likely false positives due to these regions not being captured efficiently leading to the variants not being called in the parent who carries it. Calling *de novo* mutants from bam files using specialised software to take into account read depth in the three samples would likely be a way to reduce these false positives.

7.3.4 Homozygote variants

Six samples with homozygous rare variants at microRNA loci are found excluding those on the X chromosome, they are all from the MOPD cohort, four in cases of unknown cause, one a parental control and one a case of known cause (Table 7.5). 17:79099125:G/C homozygote in parental control sample 410 is heterozygote in their affected child who has an identified disease causing mutation. Those in the four cases of unknown cause are singleton variants only seen once in these cohorts. The 3:48357921:A/G variant in sample 583 is seen as a heterozygote in

sample	microRNA	region	group	cause	chr	pos	ref	alt
411	mir-657	miRNA-5p	MOPD	control	17	79099125	G	C
454	mir-92a-1	miRNA-5p	MOPD	unknown	13	92003581	T	C
583*	mir-2115	miRNA-5p	MOPD	unknown	3	48357921	A	G
584*	mir-520d	flank-5p	MOPD	known	19	54223342	C	G
587*	mir-4524a	flank-5p	MOPD	unknown	17	67095769	C	T
608*	mir-302d	flank-3p	MOPD	unknown	4	113569157	C	G

Table 7.5: Homozygous mutations at microRNA loci, candidates to affect both copies of the microRNA. *These samples were added to the database later so have not been investigated thoroughly

one ExAC r0.3 sample of 2780 with coverage. The 17:67095769:C/T variant in sample 587 is seen as a heterozygote in one ExAC r0.3 sample of 2 with coverage. The other two variants in samples 454 and 608 are not seen in ExAC r0.3 or the 1000 Genomes Project samples. Parental genotypes are not available for these samples to identify the mode of inheritance. One of these - sample 454 - with a homozygous mutant in the 5p microRNA of miR-92a is investigated further below as this microRNA has been associated with disease relevant traits previously.

A homozygous variant in a trait associated locus

This homozygous variant of miR-92a-1 in an MOPD sample seemed particularly interesting as the miR-17∼92 cluster of microRNAs has been previously implicated in the definition of body size. De Pontual et al[193] reported germline hemizygous deletions of the miR-17∼92 microRNA cluster in three individuals with phenotypes consistent with Feingold syndrome which includes microcephaly and short stature and without mutations in MYCN - a previously established cause of Feingold syndrome. These individuals had approximately 50% expression of all six microRNAs from the cluster measured by RT-qPCR in peripheral white blood cells.

An allelic series of mutant mice with engineered deletions of combinations of microRNAs at the miR-17∼92 locus revealed specialisation and cooperation which can coexist between members of the locus with larger deletions causing skeletal abnormalities and reduced bodyweight. Deletion of miR-92 from this locus causing a ∼20% reduction in bodyweight, but this was the smallest size reduction phenotype of the targeted deletions from the locus[258].

Processing of microRNAs from this locus occurs in a complex series of steps where a progenitor microRNA expressed from this locus containing the six microRNA hairpins is dynamically regulated due to the presence of a repression domain at the base of the miR-92a-1 hairpin controlling the folding of the RNA molecule inhibiting further processing by the microprocessor[259].

The homozygous mutant at miR-92a-1 in the MOPD patient is a T>C change at the 4th position of the 5p microRNA within the seed region. Expression of the 3p microRNA from this hairpin is two orders of magnitude higher than the 5p microRNA in LCLs (Using expression data used

in section 6), though both are highly expressed relative to other microRNAs with miR-92a-1-3p being the 5th most highly expressed microRNA and miR-92a-1-5p being 160th, of 711 microRNAs expressed in >50% of samples LCLs.

Were this miR-92a-1 mutant to be causal for MOPD this could be due to the altered 5p microRNA seed sequence, altered processing of the miR-92a-1 hairpin by the microprocessor, or altered processing of the miR-17∼92 cluster.

The observation of a homozygous novel variant is surprising and suggests the variation may be hemizygous, opposite a deletion (implying Feingold syndrome). Alternatively, it may indicate consanguinity or more distant identity by decent. Examining the overall level of homozygosity in this patient revealed that they are consanguineous, with 15-20% of the genome in large (>1.5mb) runs of homozygosity depending on the tool used to calculate this. Discrepancies between methods for measuring runs of homozygosity may be expected as coverage in exome data is highly non-uniform. Using sliding windows across the genome of 1Mb requiring 100 SNPs in a window PLINK[260] finds 18% of the genome to be in runs of homozygosity. H3M2[220], a tool designed to identify regions of homozygosity from whole-exome data using a hidden Markov model, finds 16% of the genome in large runs of homozygosity. These percent homozygous scores would be consistent with a first cousin marriage. The miR-17∼92 cluster on chromosome 13 is identified as within a run of homozygosity by both of these tools (Fig 7.4).

As this patient has multiple large runs of homozygosity other variants may be causative for the MOPD phenotype, although none have been identified in known disease causing genes. This patient has 121 nonsense or missense variants which pass quality filtering, 12 homozygous and 109 heterozygous. These variants can be examined for their predictive effect on the target protein and how that protein is connected to known disease linked pathways. Examining just the homozygous or compound heterozygous variants 13 genes contain variants with varying levels of uniqueness, effect on the protein and likelihood of protein to be linked to MOPD. A novel variant in FGFR2 - involved in embryonic development of blood vessels and bone - would seem the most likely of these to be causative given its connection to embryogenesis, although the variant is predicted as tolerated/benign by sift/polyphen. None of these other homozygous mutants represents a unique variant predicted to be deleterious to protein function with a protein connected to DNA replication / DNA repair.

7.3.5 Compound heterozygous variants

Another expected mode of inheritance for the rare Mendelian disease cohorts is multiple rare variants at a locus affecting both copies of a gene. 37 samples from the database have multiple rare variants at the same microRNA locus (table 7.6). Many of these will be clusters of variants in LD which are segregating together in the population, present on only one copy of a locus. Differentiating these cases from genuine compound heterozygosity would require examining the parental genotypes or examining reads spanning these loci in the original bam files. An

group	cause	samples
Eye		7
Micro		4
MOPD	known	5
MOPD	control	11
MOPD	unknown	8
CRC		2

Table 7.6: Sample counts for potential compound heterozygotes, microRNA loci with multiple mutations in the same sample where potentially both copies of a microRNA are affected.

analysis not possible using only the variant database here.

7.3.6 microRNA loci with multiple cohort specific rare variants

Here I perform a burden analysis of rare variants at microRNA loci identifying loci with rare variants in multiple individuals from the same disease cohort, a key method of identifying disease causing genes[261]. I contrast these with rare variants in individuals from other cohorts in the database having expected coverage for each microRNA. The expected coverage of microRNAs for each individual is based on the capture platform used in their exome sequencing (section 7.3.1). Fisher's exact tests are performed for each microRNA locus to identify those with significant enrichment in one disease cohort above a background rate using variants at all other loci. One confounding factor in this analyses is that each disease cohort is recruited from populations of differing geographic location, causing population stratification which may lead to false positive enrichments where population specific variants exist. While the majority these variants are heterozygote and expected mode of inheritance in the Mendelian disease cohorts is often recessive rare variants, it is possible that additional mutants nearby may be missed due to low or no coverage in the exome target capture kits, or nearby segmental copy changes may be difficult to identify in exome sequencing.

MOPD

In the burden analysis comparing MOPD cases of unknown cause with controls and those of known cause two loci passed nominal significance with rare variants in only three cases of unknown origin at each loci (tables 7.7, 7.8). Both of these were not robust to Bonferroni multiple testing correction.

miR-758 had rare variants in three MOPD cases with no known disease causing mutation, each heterozygous mutants in the 5p flanking region with good read coverage. Two of these variants are also seen in ExAC r0.3 and the variant in sample 585 is also seen in two samples from the Eye cohort. Two other variants at this locus are also present in the database also seen in ExAC r0.3 in the 5p mature microRNA in a colorectal cancer sample and in the terminal loop in a sample from the eye cohort (table 7.8).

microRNA	cohort	samples	coverage	others_samples	others_coverage	pvalue
mir-758	MOPD	3	110	0	222	0.038
mir-4289	MOPD	3	96	0	191	0.039

Table 7.7: microRNA loci enriched for rare variants in MOPD cases of unknown cause. Numbers indicate the counts forming a contingency table used in a Fisher's exact test for samples with coverage in the named microRNA and all other microRNAs in this cohort compared to others, pvalue shown

microRNA	chr	pos	region	maf	group	samples
mir-4289	9	91360766	miRNA-3p	0.00539	MOPD	2
mir-4289	9	91360814	flank-5p	0.00030	MOPD	1
mir-758	14	101492341	flank-5p	0.00003	MOPD	1
mir-758	14	101492342	flank-5p	0.00040	Eye	2
mir-758	14	101492342	flank-5p	0.00040	MOPD	1
mir-758	14	101492362	flank-5p		MOPD	1
mir-758	14	101492383	miRNA-5p	0.00002	CRC	1
mir-758	14	101492407	loop	0.00023	Eye	1

Table 7.8: All variants at microRNA loci selected as enriched for MOPD variants

miR-4289 had rare variants in three MOPD cases of unknown cause: at one site in the 3p mature microRNA seen in two cases and a site in the 5p flank seen in one case. These are not seen in other samples in the database, but are both seen in ExAC r0.3.

Other cohorts

For the remaining cohorts where well matched control populations were not available each cohort was contrasted with all other cohorts when testing for enrichment of rare mutations at microRNA loci. After applying a Bonferroni correction for multiple testing only three microRNA loci had significant enrichments, all in the micro syndrome cohort (table 7.9).

In miR-147a where two variants were each seen in the same four samples with micro syndrome, not present in ExAC r0.3. Also present in this locus was a variant seen in two other cohorts in the database as well as in ExAC r0.3 (table 7.10).

Variants in miR-365a and miR-6874 were also seen in three and four samples with micro syndrome respectively, not seen in other cohorts in the database but both seen in ExAC r0.3.

These three significant enrichments after multiple testing correction in the Micro syndrome cohort are due to five samples with the same rare variants present amongst three or four of the samples at these loci. These signals may be due to a population specific effect from these samples, which have the same combinations of rare variants, not seen in other cohorts which are predicted to have coverage at these loci.

Within the colorectal cancer cohort there were four microRNA loci with nominal significance for enrichment of rare variants (table 7.11), none significant after multiple testing correction.

microRNA	cohort	samples	coverage	other_samples	others_coverage	pvalue
mir-147a	Micro	4	35	2	1260	9.940E-06
mir-6874	Micro	4	35	0	573	1.421E-05
mir-365a	Micro	3	35	0	1260	2.320E-05

Table 7.9: microRNA loci enriched for rare variants in the Micro syndrome cohort. Numbers indicate the counts forming a contingency table used in a Fisher's exact test for samples with coverage in the named microRNA and all other microRNAs in this cohort compared to others, pvalue shown

microRNA	chr	pos	region	maf	group	samples
mir-147a	9	123007254	flank-3p		Micro	4
mir-147a	9	123007266	miRNA-3p		Micro	4
mir-147a	9	123007313	miRNA-5p	0.001997	MOPD	1
mir-147a	9	123007313	miRNA-5p	0.001997	CRC	1
mir-365a	16	14403139	flank-5p	0.000017	Micro	3
mir-6874	7	5751517	miRNA-5p	0.000008	Micro	4

Table 7.10: All variants at microRNA loci enriched for mutations in the micro syndrome cohort

microRNA	cohort	samples	coverage	other_samples	others_coverage	pvalue
mir-96	CRC	19	672	3	623	0.001
mir-499a	CRC	6	672	0	623	0.032
mir-499b	CRC	6	672	0	623	0.032
mir-3614	CRC	7	672	0	424	0.048

Table 7.11: microRNA loci enriched for rare variants in the CRC cohort. Numbers indicate the counts forming a contingency table used in a Fisher's exact test for samples with coverage in the named microRNA and all other microRNAs in this cohort compared to others, pvalue shown

microRNA	chr	pos	region	maf	group	samples
mir-3614	17	54968700	miRNA-5p	0.0023	CRC	7
mir-499a	20	33578188	flank-5p	0.0004	CRC	8
mir-499a	20	33578255	miRNA-3p	0.0008	CRC	4
mir-96	7	129414568	loop	0.0081	Eye	2
mir-96	7	129414568	loop	0.0081	CRC	12
mir-96	7	129414574	loop	0.0058	MOPD	1
mir-96	7	129414574	loop	0.0058	CRC	7

Table 7.12: All variants at microRNA loci enriched for CRC variants

Two variants in miR-96 present in 7 and 12 samples from the CRC cohort were also seen in 1 and 2 samples from other cohorts and were present in ExAC r0.3.

Within the cranio-facial developmental disease cohort (Eye) 19 microRNA loci passed nominal significance (table 7.13), with none robust to Bonferroni multiple testing correction.

microRNA	cohort	samples	coverage	others_samples	others_coverage	pvalue
mir-770	Eye	4	241	0	1054	0.001
mir-323a	Eye	3	241	0	1054	0.007
mir-5196	Eye	4	241	2	1054	0.013
mir-4285	Eye	3	98	0	311	0.014
mir-4746	Eye	6	241	1	367	0.019
mir-26b	Eye	3	241	1	1054	0.023
mir-6721	Eye	3	241	1	1054	0.023
mir-181d	Eye	3	241	1	1054	0.023
mir-509-3	Eye	4	241	2	878	0.023
mir-1236	Eye	4	241	0	367	0.025
mir-1228	Eye	4	241	0	367	0.025
mir-874	Eye	2	241	0	1054	0.035
mir-34c	Eye	2	241	0	1054	0.035
mir-199b	Eye	2	241	0	1054	0.035
mir-635	Eye	2	241	0	1054	0.035
mir-99b	Eye	2	241	0	1054	0.035
mir-128-2	Eye	2	241	0	1054	0.035
mir-516a-1	Eye	2	241	0	878	0.047
mir-3909	Eye	3	98	1	311	0.047

Table 7.13: microRNA loci enriched for rare variants in the Eye cohort. Numbers indicate the counts forming a contingency table used in a Fisher's exact test for samples with coverage in the named microRNA and all other microRNAs in this cohort compared to others, pvalue shown

7.3.7 Mutations in microRNAs which target known disease genes

microRNA targets within 3'UTRs predicted by TargetScan v7 were analysed to see if previously identified known causal genes are targeted by microRNAs which may have their effects altered in these cohorts through mutation, either directly changing the sequence of the mature microRNA or indirectly altering the expression of the microRNA. For each cohort a Fisher's exact test was performed contrasting microRNAs predicted to target disease implicated genes with all other microRNAs, between the selected cohort and all other samples in the database.

Genes were selected where mutations were known to cause disease using published reviews and internal up to date lists.

Genes containing mutations identified as causal in MOPD have been identified in a variety of pathways, I have selected 45 based on published results: DNA damage response: ATR[262], BLM[263], DNA2[264], LIG4[265], MRE11A[266], PNKP[267], XRCC4[264]; Centrasomal proteins: ASPM[268], PCNT[269], CEP152[270], CENPF[271], CENPJ[272], CDK5RAP2[273], CPAP[272], NCAPD2[274, 275], NCAPD3[274]; The prereplicative complex: ORC1[276, 277], ORC4[278, 277], ORC6[277], CDC6[277], CDC45[279], CDT1[277]; Splicing machinery: RBM10[267], RNU4[280], U4atac[280]; And others: ANKLE2[281], BCOR[282], CREBBP[283], DYRK1A[284], ERCC6[285], ESCO2[286], FOXG1[287], KIF11[288], MED12[289], NDUFB3[290], NSUN2[291], PIK3R1[292], PLK4[293], SIL1[294], SMARCAL1[295], SRCAP[296], TUBGCP6[297], VPS13B[298].

For the eye developmental disease cohort 11 genes were selected as linked to eye development based on published reviews[299, 300]: CHX10[301], RAX[302], PAX6[303], BCOR[282], SOX2[304], OTX2[305], RAB3GAP[306], STRA6[307], BMP4[308], SMOC1[309], and FOXE3[310].

Five genes linked with micro syndrome were examined for mutations in microRNAs predicted to target them: RAB3GAP1[306], RAB3GAP2[306], RAB18[311], TBC1D20[312], and ITPA[313].

Colorectal cancer implicated genes used were also used to examine mutations in microRNAs predicted to target them: APC[314], TP53[315], KRAS[316], PIK3CA[317], FBXW7[318], SMAD4[319], TCF7L2[320], NRAS[316], BRAF[321], SMAD2[322], PCBP1[323].

Results for the Fisher's exact tests are shown in table 7.14, no cohort was significantly enriched for rare variants in microRNAs targeting disease implicated genes. Due to the relatively small sample sizes of the cohorts in this analysis the power to detect a signal here is low. Where larger sample sizes are available this analysis may be more likely to detect a signal.

group	target_group	non_target_group	target_others	non_target_others	pvalue
MOPD	20	34	34	65	0.86
Eye	24	48	103	244	0.57
Micro	2	8	118	284	0.73
CRC	116	137	123	155	0.73

Table 7.14: Fisher's exact tests for enrichment of samples with rare mutations in microRNAs targeting disease implicated genes for each cohort

7.4 Chapter summary

microRNA loci are non-protein-coding regions of the genome which are frequently overlooked in disease genetics and variant annotation pipelines as predicting the effect of variants on microRNA function is an ongoing problem.

As the amount of whole genome sequencing of patients increases this will become a more pressing issue, and as I have shown many microRNA loci are captured by existing exome capture kits frequently used in current patient sequencing studies.

As with common variation (section 6.3.1) rare variants are distributed across microRNA loci. Distinguishing pathogenic variants from non-pathogenic will require analyses such as those demonstrated here identifying loci enriched for variants in large disease cohorts. Use of microRNA targeting databases can also identify those microRNAs which are more likely to be functionally relevant to disease.

I have performed these variant prioritisation analyses in both rare Mendelian disorders and in the common late onset disease colorectal cancer:

In MOPD samples with no known causal mutation I have identified several homozygous variants at microRNA loci, investigating one in a patient were this variant is in miR-92a, a member of a microRNA cluster where deletion studies have shown skeletal and body size malformations. While consanguinity in this sample and potentially others will cause segments of the genome to be in runs of homozygosity, increasing the number of potential disease causing variants, where standard screening methods in protein coding genes have so far failed to identify a causal mutation microRNA loci are interesting candidates. Screening for additional mutations at the miR-92a locus is currently being performed by our collaborators in larger cohorts of patients with similar phenotypes which have not undergone this exome sequencing as additional cases with mutations at his locus would be compelling evidence for causality and further investigation of these variants.

Within the MOPD samples where trios had been sequenced I have identified four putative *de novo* microRNA mutations at microRNA loci, however due to the exome capture method and variant information available in the database used this analysis is enriched for regions which are poorly captured, and it is likely that in the parents there is not sufficient read depth to call these variants.

I have performed burden analyses examining individual microRNA loci and groups of microRNA loci where the mature microRNA is predicted to target a disease implicated gene, contrasting each cohort with the other cohorts in the database. Grouping microRNA loci by their predicted targets there was so significant enrichment for rare variants in cohorts which target disease implicated genes. This may not be surprising due to both the low sample sizes of the cohorts here, and the high false positive rate for microRNA target prediction algorithms.

Examining individual loci in a burden analysis for rare variants identified three significant loci robust to Bonferroni correction in micro syndrome, and nominally significant variants in the MOPD, cranio-facial malformation, and colorectal cancer cohorts. Where well matched control populations are not available this analysis is likely to be enriched for population specific rare variants due to founder effects. This may be the case for the nominally significant results in the colorectal cancer and cranio-facial malformations cohorts, as well as the significant results in the micro syndrome cohort. Where a well matched control was available in contrasting MOPD patients with and without identified disease causing mutations two loci were only nominally significant each with rare variants in three samples.

7.4.1 Future Directions

As cohort sizes become larger in exome and increasingly in whole-genome sequencing studies additional attention will be focused on non protein-coding regions of the genome when standard variant screening procedures do not identify causal mutations and when the heritability of phenotypes cannot be explained. In investigating microRNA loci the complementary approaches taken in this chapter can form the basis for identifying additional candidate disease causing mutations.

Additional screening is being performed at the miR-92a locus where a homozygous variant in an MOPD sample with no known disease causing mutation was identified as additional similar cases with variants at this locus would provide compelling evidence to investigate the causality of these variants.

The approaches taken here may also be applied to larger exome sequencing databases such as the ExAC database[209], where although much phenotypic data is not available the power to detect enrichments would be greater due to the larger sample sizes.

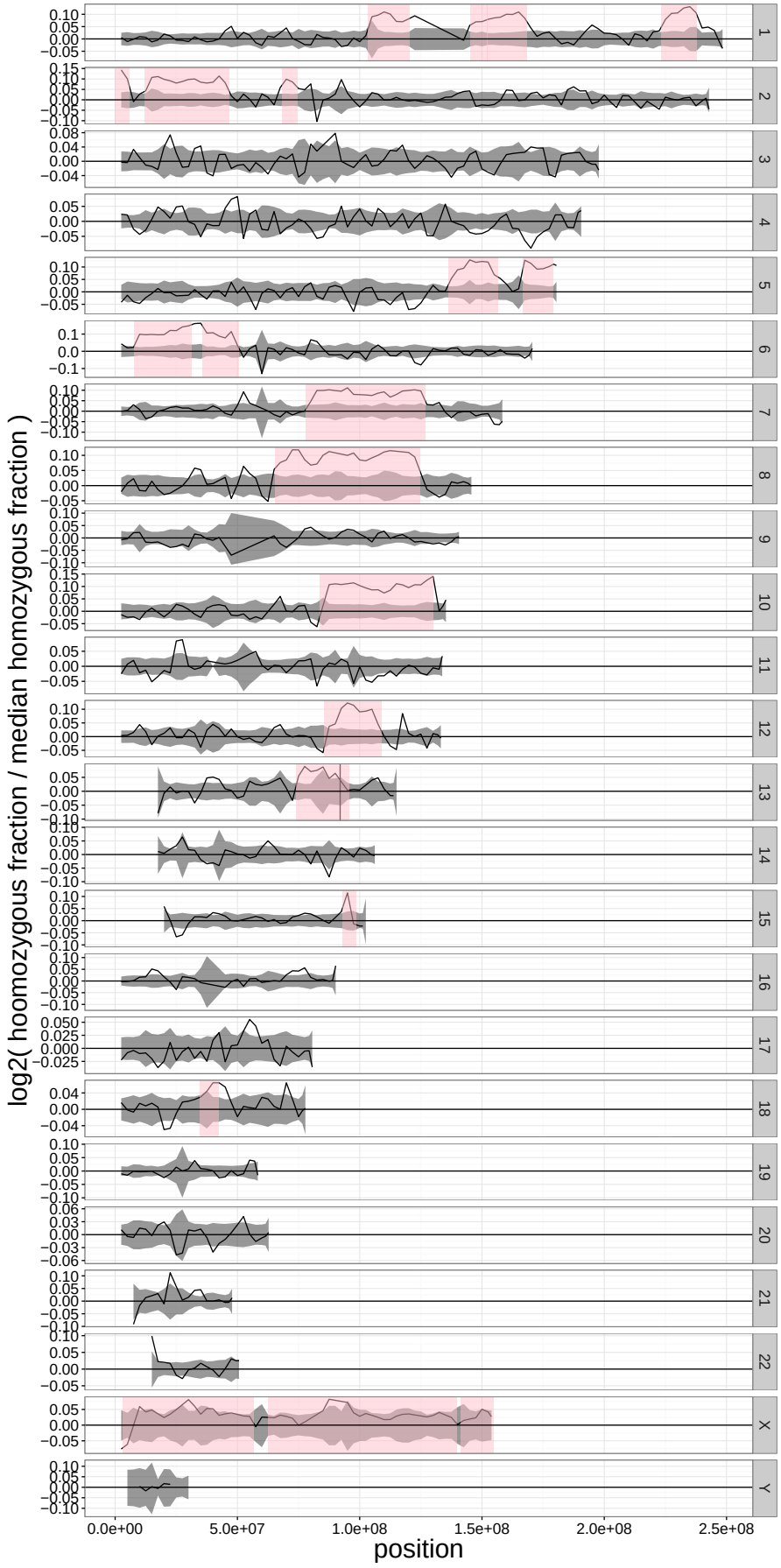


Figure 7.4: Homozygosity across the genome in MOPD patient. (x-axis) chromosome position. (y-axis) sliding window homozygosity for patient (black line) other samples from the same capture platform interquartile range (grey ribbon). Predicted runs of homozygosity by H3M2 (pink). miR-17∼92 cluster (vertical black line in chromosome 13)

Chapter 8

Discussion

The main aims of this project were the use of computational resources to generate and test hypotheses relating to microRNA production and functions: Firstly to examine the diversity of microRNA targeting using CLASH data evaluating the evidence for competitive endogenous RNAs (ceRNAs) and non-canonical microRNA functions. Secondly to examine the effect of genetic variation on microRNA processing using publicly available data, and the effect of genetic variation at microRNA loci in disease.

8.1 A pipeline for the analysis of CLASH data

Addressing the first of my project aims required the development of a pipeline for the analysis of data from the cross-linking, ligation and sequencing of hybrids (CLASH) protocol[100]. When applied to Argonaute this protocol produces a subset of sequencing reads which are part microRNA and part target sequence which the microRNA was paired with, as well as reads which are simply the target sequence like that of a standard CLIP experiment.

In collaboration I have improved upon and published a pipeline for the analysis of CLASH data, providing improvements to the run-time which out-performs an alternative method in number and quality of hybrids identified. This pipeline is available online freely under GNU license[202].

Building on this work I have also developed a pipeline for the analysis of CLASH data aligning to a genomic database. Parameters for this pipeline were obtained from a parameter sweep to maximise the number and quality of quality of hybrids, measuring quality through how well microRNAs base-pair with their identified targets as no true positive dataset was known. This pipeline has been applied for the identification of microRNA targets to address my first project aim in chapters 4 and 5, and has also been used in genomic breakpoint mapping in cancer within the IGMM[230].

8.2 The genome-wide diversity in microRNA targeting

microRNAs are known to base-pair with a variety of transcript types, however they have been mainly studied in the context of their canonical base-pairing with protein-coding genes 3'UTRs leading to translational repression. The computational prediction of microRNA targets is a well studied problem in computational biology with a variety of prediction algorithms published with many false positive predictions expected due to the lack of overlap between algorithms (section 1.4.2). Many current algorithms also limit the prediction of microRNA targets to within the 3'UTRs of protein coding genes.

The CLASH dataset[101] used in chapters 4 and 5 represents the first high throughput assay of direct microRNA–target identification, previously examined for protein coding genes identifying different classes of microRNA targeting - primarily seed region, compensatory 3' binding, non-canonical. Further examination of the clash dataset by the Bartel lab[96] suggested that those non-canonical targets did not mediate repression of the target transcript, despite microRNA binding.

These non-repressive interactions, as well as all interactions outside of protein-coding genes 3'UTRs will contribute to the pool of available microRNA targets and may also have additional functions through the protein-protein interactions of Argonaute.

8.2.1 Competitive RNA activity

All targets of a microRNA will affect the microRNA:target ratio which in turn may affect the level of repression due to the sequestration of microRNAs. Several cases of competitive endogenous RNA transcripts have been described including pseudogenes, lincRNAs and circularised exons formed by back-splicing.

Chapter 4 examined these transcripts for targets in the CLASH dataset, few targets were identified with more microRNA targeting from a single microRNA than expected looking at the overall distribution of microRNA targets.

While the previously identified miR-7 association with CDR1as was seen as the most significant association between microRNA and circular RNA no other association with predicted circular RNAs were robust to multiple testing correction.

Other transcripts with significant enrichment for multiple target sites of a specific microRNA were TP73-AS1 with miR-125 and MAP3K1 with two members of the Let-7 family of microRNAs, making them candidates for genuine microRNA targeting or competitive RNAs acting to affect the level of free microRNAs available for repression.

Using RNA sequencing data to measure transcript abundance, a linear relationship could be seen between transcript abundance and the level of CLASH binding to a transcript, which was not present for intronic sequences. Modelling these interactions on a per-microRNA basis

allowed the identification of transcripts with higher levels of binding than might be expected if effects were purely stochastic, representing potential genuine canonical targets or competitive RNA species.

Examining spatial clusters of all microRNA binding sites in chapter 5 identified regions of transcripts which have multiple independent binding sites, biased towards highly expressed transcripts. Examining Argonaute targets identified through the more dense dataset of all CLASH reads provides clusters less biased towards highly expressed genes and may be candidates sources of competitive RNAs.

8.2.2 non-canonical functions

Several papers have described the targeting of microRNAs or siRNAs to transcription start sites or splice sites having an effect on those processes, usually through the interaction of Argonaute with chromatin remodelling proteins or splicing factors.

microRNAs targeting transcription start sites were examined using the HEK293 cell type specific FANTOM5 TSS derived from CAGE-seq[207]. This identified two microRNAs miR-1307 and miR-935 each with target sites within 50bp of nine different TSS. While significantly more than expected by chance in each case these microRNAs had a greater number of non-TSS targets sites.

Similarly seven microRNAs were seen to target sites within 50bp of splicing donor or acceptor sites more often than would be expected by chance, in each case representing between 50% and 75% of targets, with miR-320 also having a specific peak of binding sites upstream of the splicing acceptor site.

8.2.3 Future work

As the effect of microRNA repression, dependent on microRNA expression levels and target site abundance, is cell-type specific. It is plausible that within specific cells or tissues transcripts may be expressed which will compete effectively for microRNA targeting at a sufficient level to affect their repression activity.

Within the dataset examined there are few transcripts with more target sites for specific microRNAs than would be expected by chance. Studies in additional cell types identifying cell-type specific transcripts and using target site prediction software, or direct CLIP/CLASH assays may uncover these.

The activity of microRNAs and Argonaute affecting additional non-canonical processes including transcription and splicing remains a controversial subject. In this dataset several microRNAs were present targeting transcription start sites or splice sites more often than would be expected by chance.

Functional testing of these enrichments could be performed through transfection of cells with microRNA mimics or microRNA inhibitors, assays of gene expression or splicing with site specific qPCR or genome-wide via microarrays or RNA-seq.

8.3 The effect of genetic variation in microRNA processing

The processing of microRNA hairpins from within primary microRNA transcripts to produce mature microRNAs is a multi-stage process occurring first in the nucleus where precursor-microRNAs are excised from RNA hairpins by DROSHA/DGCR8. Pre-microRNAs are then exported to the cytoplasm where further processing by DICER produces single stranded mature microRNAs to be loaded into Argonaute proteins to form the RNA-induced silencing complex (RISC).

Each stage of this production process may be dependant on the binding of proteins to sequence motifs, or the presence of RNA structural features, each of which may be affected by sequence variants.

Several studies have used DNA markers or genome-sequences with measures of small RNA abundance to identify microRNA eQTLs, however they have not investigated the mechanisms of these effects. Using data from the largest such study of small RNA sequencing in 465 samples sequenced as part of the 1000 Genomes Project[186, 204] I have identified rare and common variants which are eQTLs for the microRNAs they are contained within.

As whole genome sequencing has been performed it can be investigated whether these variants are the functional variants or whether other nearby variants may be responsible for this effect.

Several of these variants have also been associated with cancers and blood lipid traits. Additionally one candidate processing variant for miR-146a displayed a reciprocal response between 5p and 3p microRNA expression whereas all other microRNAs where a 3p and 5p microRNA were expressed displayed the same direction of change, suggesting a different mechanism in this case.

To aid in the prediction of candidate variants derived allele frequency tests were used to examine the evidence for selection, and Fisher's exact tests for eQTL activity of variants within different microRNA hairpin components. Derived allele frequency tests demonstrated that the microRNA loci considered while having more selective constraint than the genome-wide background they had less selective constraint than protein-coding codons as a whole. Examining individual components of microRNA hairpins the number of variants to be examined became relatively small leading to large error bars, with the error bars of each component overlapping, and 5p microRNAs showing evidence for positive selection with an odds ratio less than 1, although not significantly. Similarly the odds ratio of variants of being a microRNA eQTL was significantly higher for those within primary-microRNA stem loops than in the flanking sequence. While examining microRNA hairpin components significant

odds ratios were seen for the internal components of the stem with relatively large error bars due to the small numbers of variants.

8.3.1 Future work

A grant application has been submitted to the European Research Council to follow up the candidate microRNA processing defects with a battery of experimental approaches. Where processing variants are confirmed they will be investigated further such as in the case of the novel miR-30c-1 variant where SHAPE, hydroxyl radical cleavage footprinting, and protein binding assays were used to examine the effect of a variant on RNA structure of the microRNA hairpin and protein binding[199] (section 1.7.4).

Knowledge about which variants are likely to affect microRNA processing can then be applied in computational tools for predicting variant consequences for the annotation of variants in trait and disease sequencing studies.

Combining these molecular phenotypes (microRNA expression level), experimental testing of microRNA processing and the measurement of downstream target expression by RNA-seq could provide rapid mechanistic insight into the impact of common non-coding variation on human traits.

Larger whole genome sequence datasets would provide additional variants for derived allele frequency tests to aid in examining evidence for selection, also useful in the prioritisation of variants.

8.4 The effect of genetic variation at microRNA loci in disease

Several cases of genetic variants at microRNA loci have been seen to cause disease, however variants at microRNA loci are often overlooked in variant annotation and pathogenicity prediction tools.

In collaboration with clinically focussed research groups at the MRC HGU (including the groups of A. Jackson, D.Fitzpatrick and M. Dunlop) I have used several in-house exome sequencing studies totalling 1295 samples in a pilot study for the screening of variants at microRNA loci to identify candidate pathogenic variants.

The use of these cohorts had the advantage that additional phenotypic information is available and where any candidate variants are identified they can be followed up on in collaboration with the disease genetic group.

Using disease models relevant for each disease, e.g. (1) rare recessive variants seen as homozygotes or compound heterozygotes in the rare MOPD and cranio-facial developmental

diseases. (2) Many common and rare variants of low penetrance contributing to disease likelihood in colorectal cancer. I have screened variants for their potential pathogenicity. Due to the small size of these datasets, the power to detect pathogenicity of variants through a burden analysis where genes are mutated in several individuals is low. Particularly given the lack of well matched controls, leading to population stratification between sample groups in which population specific and disease causing variants are difficult to distinguish.

This study identified a proband that is homozygous for a private variant in miR-92a with microcephalic osteodysplasic primordial dwarfism (MOPD). This sample from an MOPD patient with no identified pathogenic variant in a disease relevant gene. This proband does not harbour candidate causal variants in genes known to be associated with primordial dwarfism, and hemizogysity for this locus leads to growth deficits in both humans and mice.

8.4.1 Future work

This study has shown that a majority of microRNA loci, and almost all high confidence microRNA loci are captured in current exome sequencing capture kits. Screening of additional exome sequencing datasets using the procedures used here would provide additional candidate disease causing variants.

As the use of exome and whole-genome sequencing accelerates need for effective variant screening pipelines will increase, particularly in the non-protein-coding regions of the genome where effective tools are currently lacking.

Further elucidation of rules governing the effect of variants in microRNA processing and targeting, through methods such as those applied this project, will be required to build those tools.

Working in collaboration with the disease groups at the IGMM has meant that additional sample information and follow up studies are easily available. Examining large cohorts such as the ExAC dataset while potentially providing additional candidate variants would be difficult to interpret due to the lack of phenotypic information about samples due to data sharing restrictions.

The identified candidate processing variant in miR-92a is being followed up with screening of sample panels from patients of similar phenotypes which have not been exome sequenced, and testing in microRNA processing assays. Candidate processing defects such as this represent a model for the future characterisation of transcribed non-coding nucleotide sequence variants.

References

[1] Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75:843–854.

[2] Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 75:855–862.

[3] Reinhart BJ et al. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 403:901–906.

[4] Pasquinelli aE et al. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 408:86–89.

[5] Lee RC, Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)* 294:862–4.

[6] Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) 2001 An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans.pdf. *Science (New York, N.Y.)* 294:858–62.

[7] Grad Y et al. (2003) Computational and experimental identification of C. elegans microRNAs. *Molecular Cell* 11:1253–1263.

[8] Landgraf P et al. (2007) A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* 129:1401–1414.

[9] Creighton CJ, Reid JG, Gunaratne PH (2009) Expression profiling of microRNAs by deep sequencing. *Briefings in Bioinformatics* 10:490–497.

[10] Liao JY et al. (2010) Deep sequencing of human nuclear and cytoplasmic small RNAS reveals an unexpectedly complex subcellular distribution of mirnas and tRNA 3 trailers. *PLoS ONE* 5:e10563.

[11] Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ (2008) miRBase: Tools for microRNA genomics. *Nucleic Acids Research* 36:D154–8.

[12] Lagos-quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes Coding for RNAs of Small expressed RNAs. *Science* 294:853–858.

[13] Elbashir SM et al. (2001) Duplexes of 21 nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411:494–498.

[14] Fire A et al. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 391:806–811.

[15] Moffat J et al. (2006) A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. *Cell* 124:1283–1298.

[16] Bumcrot D, Manoharan M, Koteliansky V, Sah DWY (2006) RNAi therapeutics: a potential new class of pharmaceutical drugs. *Nature Chemical Biology* 2:711–719.

[17] Cox DN et al. (1998) A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes and Development* 12:3715–3727.

[18] Harris aN, Macdonald PM (2001) Aubergine encodes a Drosophila polar granule component required for pole cell formation and related to eIF2C. *Development (Cambridge, England)* 128:2823–2832.

[19] Deng W, Lin H (2002) Miwi, a Murine Homolog of Piwi, Encodes a Cytoplasmic Protein Essential for Spermatogenesis. *Developmental Cell* 2:819–830.

[20] Kuramochi-Miyagawa S et al. (2004) Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development (Cambridge, England)* 131:839–849.

[21] Aravin A et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442:203–207.

[22] Girard A, Sachidanandam R, Hannon GJ, Carmell Ma (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442:199–202.

[23] Lau N et al. (2006) Characterlization of the piRNA Complex from Rat Testes. *Science* 313:363–367.

[24] Grivna ST, Pyhtila B, Lin H (2006) MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proceedings of the National Academy of Sciences* 103:13415–13420.

[25] Prestayko AW, Tonato M, Busch H (1970) Low molecular weight RNA associated with 28 s nucleolar RNA. *Journal of Molecular Biology* 47:505–515.

[26] Kiss-László Z, Henry Y, Bachellerie JP, Caizergues-Ferrer M, Kiss T (1996) Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs. *Cell* 85:1077–1088.

[27] Tabara H et al. (1999) The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell* 99:123–132.

[28] Grishok A (2000) Genetic Requirements for Inheritance of RNAi in C. elegans. *Science* 287:2494–2497.

[29] Bernstein E, Caudy aa, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409:363–366.

[30] Tabara H, Yigit E, Siomi H, Mello CC (2002) The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-Box helicase to direct RNAi in C. elegans. *Cell* 109:861–871.

[31] Sasaki T, Shiohama A, Minoshima S, Shimizu N (2003) Identification of eight members of the Argonaute family in the human genome. *Genomics* 82:323–330.

[32] Wang D et al. (2012) Quantitative functions of argonaute proteins in mammalian development. *Genes and Development* 26:693–704.

[33] Su H, Trombly MI, Chen J, Wang X (2009) Essential and overlapping functions for mammalian argonautes in microRNA silencing. *Genes and Development* 23:304–317.

[34] Meister G et al. (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular Cell* 15:185–197.

[35] Liu J et al. (2001) A slice of the action. *Engineer* 290:22.

[36] Burroughs AM et al. (2011) Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA biology* 8:158–177.

[37] Denli AM, Tops BBJ, Plasterk RHa, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature* 432:231–5.

[38] Gregory RI et al. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature* 432:235–240.

[39] Lee Y et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425:415–419.

[40] Landthaler M, Yalcin A, Tuschl T (2004) The human DiGeorge syndrome critical region gene 8 and its D. melanogaster homolog are required for miRNA biogenesis. *Current Biology* 14:2162–2167.

[41] Chiang HR et al. (2010) Mammalian microRNAs: Experimental evaluation of novel and previously annotated genes. *Genes and Development* 24:992–1009.

[42] Chen X (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science (New York, N.Y.)* 303:2022–2025.

[43] Han J et al. (2006) Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell* 125:887–901.

[44] Ha M, Kim VN (2014) Regulation of microRNA biogenesis. *Nature reviews. Molecular cell biology* 15:509–524.

[45] Auyeung VC, Ulitsky I, McGeary SE, Bartel DP (2013) Beyond secondary structure: Primary-sequence determinants license Pri-miRNA hairpins for processing. *Cell* 152:844–858.

[46] Fang W, Bartel DP (2015) The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. *Molecular Cell* 60:131–145.

[47] Alarcón CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF (2015) N6-methyladenosine marks primary microRNAs for processing. *Nature* 519:482–5.

[48] Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes and Development* 17:3011–3016.

[49] Bohnsack MT, Czaplinski K, Gorlich D (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA (New York, N.Y.)* 10:185–91.

[50] Hutvagner G et al. (2010) A Cellular Function for the RNA-Interference Temporal RNA Small let-7 Enzyme Dicer in the Maturation of the . *Science* 293:1–6.

[51] Schwarz DS et al. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115:199–208.

[52] Chendrimada TP et al. (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 436:740–4.

[53] Johnston M, Geoffroy MC, Sobala A, Hay R, Hutvagner G (2010) HSP90 protein stabilizes unloaded argonaute complexes and microscopic P-bodies in human cells. *Molecular biology of the cell* 21:1462–9.

[54] Iwasaki S et al. (2010) Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Molecular Cell* 39:292–299.

[55] Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115:209–216.

[56] Winter J, Diederichs S (2013) Argonaute-3 activates the let-7a passenger strand microRNA. *RNA biology* 10:1631–43.

[57] Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. *Cell* 130:89–100.

[58] Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC (2007) Mammalian Mirtron Genes. *Molecular Cell* 28:328–336.

[59] Taft RJ et al. (2009) Tiny RNAs associated with transcription start sites in animals. *Nature Genetics* 41:572–578.

[60] Zamudio JR, Kelly TJ, Sharp PA (2014) Argonaute-bound small RNAs from promoter-proximal RNA polymerase II. *Cell* 156:920–934.

[61] Rybak-Wolf A et al. (2014) A variety of dicer substrates in human and C. elegans. *Cell* 159:1153–1167.

[62] Pane A, Wehr K, Schüpbach T (2007) zucchini and squash Encode Two Putative Nucleases Required for rasiRNA Production in the Drosophila Germline. *Developmental Cell* 12:851–862.

[63] Olivieri D, Senti KA, Subramanian S, Sachidanandam R, Brennecke J (2012) The Cochaperone Shutdown Defines a Group of Biogenesis Factors Essential for All piRNA Populations in Drosophila. *Molecular Cell* 47:954–969.

[64] Saito K et al. (2007) Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends. *Genes & development* 21:1603–8.

[65] Kawaoka S, Izumi N, Katsuma S, Tomari Y (2011) 3' End Formation of PIWI-Interacting RNAs In Vitro. *Molecular Cell* 43:1015–1022.

[66] Brennecke J et al. (2007) Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* 128:1089–1103.

[67] Gunawardane LS et al. (2007) A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science* 315:1587–1590.

[68] Malone CD et al. (2009) Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. *Cell* 137:522–535.

[69] Rouget C et al. (2010) Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. *Nature* 467:1128–1132.

[70] Rajasethupathy P et al. (2012) A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell* 149:693–707.

[71] Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316:744–747.

[72] Aravin AA et al. (2008) A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Molecular Cell* 31:785–799.

[73] Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science* 315:241–244.

[74] Sijen T, Steiner Fa, Thijssen KL, Plasterk RHa (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science (New York, N.Y.)* 315:244–247.

[75] Ghildiyal M et al. (2008) Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science* 320:1077–81.

[76] Czech B et al. (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature* 453:798–802.

[77] Okamura K, Balla S, Martin R, Liu N, Lai EC (2008) Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. *Nat Struct Mol Biol* 15:581–590.

[78] Murchison EP et al. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453:534–8.

[79] Watanabe T et al. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453:539–543.

[80] Tollervey D, Kiss T (1997) Function and synthesis of small nucleolar RNAs. *Current Opinion in Cell Biology* 9:337–342.

[81] Kishore S et al. (2013) Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome biology* 14:R45.

[82] Ender C et al. (2008) A Human snoRNA with MicroRNA-Like Functions. *Molecular Cell* 32:519–528.

[83] Rhoades MW et al. (2002) Prediction of plant microRNA targets. *Cell* 110:513–520.

[84] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of Mammalian MicroRNA Targets. *Cell* 115:787–798.

[85] John B et al. (2004) Human MicroRNA targets. *PLoS biology* 2:e363.

[86] Krek A et al. (2005) Combinatorial microRNA target predictions. *Nature genetics* 37:495–500.

[87] Chen K, Rajewsky N (2006) Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes. *Cold Spring Harbor symposia on quantitative biology* 71:149–56.

[88] Berezikov E et al. (2006) Diversity of microRNAs in human and chimpanzee brain. *Nature genetics* 38:1375–7.

[89] Chen K, Rajewsky N (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nature genetics* 38:1452–1456.

[90] Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* 19:92–105.

[91] Hu HY et al. (2012) Evolution of the human-specific microRNA miR-941. *Nature Communications* 3:1145.

[92] Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120:15–20.

[93] Lim LP et al. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433:769–73.

[94] Liu C et al. (2013) CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Research* 41:e138.

[95] Erhard F, Dölken L, Jaskiewicz L, Zimmer R (2013) PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome biology* 14:R79.

[96] Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4:1–38.

[97] Chi SW, Zang JB, Mele A, Darnell RB (2009) Ago HITS-CLIP decodes miRNA-mRNA interaction maps. *Nature* 460:479–486.

[98] Hafner M et al. (2010) Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* 141:129–141.

[99] Leung AK et al. (2011) Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat Struct Mol Biol* 18:237–244.

[100] Kudla G, Granneman S, Hahn D, Beggs JD, Tollervey D (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. TL - 108. *Proceedings of the National Academy of Sciences of the United States of America* 108 VN -:10010–10015.

[101] Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153:654–665.

[102] Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science (New York, N.Y.)* 297:2053–2056.

[103] Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466:835–840.

[104] Eichhorn SW et al. (2014) MRNA Destabilization Is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Molecular Cell* 56:104–115.

[105] Meijer HA et al. (2013) Translational repression and eIF4A2 activity are critical for microRNA-mediated gene regulation. *Science (New York, N.Y.)* 340:82–85.

[106] Fukaya T, oki Iwakawa H, Tomari Y (2014) MicroRNAs block assembly of eIF4F translation initiation complex in drosophila. *Molecular Cell* 56:67–78.

[107] Tat TT, Maroney PA, Chamnongpol S, Coller J, Nilsen TW (2016) Cotranslational microRNA mediated messenger RNA destabilization. *eLife* 5:1–18.

[108] Landthaler M et al. (2008) Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA (New York, N.Y.)* 14:2580–96.

[109] Mathys H et al. (2014) Structural and Biochemical Insights to the Role of the CCR4-NOT Complex and DDX6 ATPase in MicroRNA Repression. *Molecular Cell* 54:751–765.

[110] Schwanhäusser B et al. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature* 455:58–63.

[111] Brodersen P et al. (2008) Widespread translational inhibition by plant miRNAs and siRNAs. TL - 320. *Science (New York, N.Y.)* 320 VN -:1185–1190.

[112] Mathonnet G et al. (2007) MicroRNA Inhibition of Translation Initiation in VItro By Targeting the Cap-Binding Complex eIF4F. *Science* 317:1764–1767.

[113] Wakiyama M, Takimoto K, Ohara O, Yokoyama S (2007) Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. *Genes and Development* 21:1857–1862.

[114] Maroney Pa, Yu Y, Fisher J, Nilsen TW (2006) Evidence that microRNAs are associated with translating messenger RNAs in human cells. *Nature structural & molecular biology* 13:1102–7.

[115] Petersen CP, Bordeleau ME, Pelletier J, Sharp PA (2006) Short RNAs repress translation after initiation in mammalian cells. *Molecular Cell* 21:533–542.

[116] Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews. Genetics* 9:102–114.

[117] Katayama S (2012) Antisense Transcription in the Mammalian Transcriptome. *Science* 1564:1564–6.

[118] Li S et al. (2013) MicroRNAs inhibit the translation of target mRNAs on the endoplasmic reticulum in arabidopsis. *Cell* 153:562–574.

[119] Meister G (2013) Argonaute proteins: functional insights and emerging roles. *Nature reviews. Genetics* 14:447–59.

[120] Huang V, Li LC (2014) Demystifying the nuclear function of Argonaute proteins. *RNA biology* 11:18–24.

[121] Volpe T (2002) Regulation oh heterochromatic silencing and histone H3 Lysine-9 by RNAi. *Science* 297:1833–1837.

[122] Yu R, Jih G, Iglesias N, Moazed D (2014) Determinants of Heterochromatic siRNA Biogenesis and Function. *Molecular Cell* 53:262–276.

[123] Wassenegger M, Heimes S, Riedel L, Sänger HL (1994) RNA-directed de novo methylation of genomic sequences in plants. *Cell* 76:567–76.

[124] Taverna SD, Coyne RS, Allis CD (2002) Methylation of histone H3 at lysine 9 targets programmed DNA elimination in Tetrahymena. *Cell* 110:701–711.

[125] Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in Tetrahymena. *Cell* 110:689–699.

[126] Brower-Toland B et al. (2007) Drosophila PIWI associates with chromatin and interacts directly with HP1a. *Genes and Development* 21:2300–2311.

[127] Sienski G, Dönertas D, Brennecke J (2012) Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* 151:964–980.

[128] Le Thomas A et al. (2013) Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes and Development* 27:390–399.

[129] Cernilogar FM et al. (2011) transcriptional regulation in Drosophila. *Nature* 480:391–395.

[130] Moshkovich N et al. (2011) RNAi-independent role for argonaute2 in CTCF/CP190 chromatin insulator function. *Genes and Development* 25:1686–1701.

[131] Rüdel S, Flatley A, Weinmann L, Kremmer E, Meister G (2008) A multifunctional human Argonaute2-specific monoclonal antibody. *RNA (New York, N.Y.)* 14:1244–1253.

[132] Robb GB, Brown KM, Khurana J, Rana TM (2005) Specific and potent RNAi in the nucleus of human cells. *Nature structural & molecular biology* 12:133–137.

[133] Hwuang H.-W., Wentzel E. MJ (2007) A Hexanucleotide Element Directs MircoRNA Nucelar Import. *Science* 315:97–101.

[134] Politz JCR, Hogan EM, Pederson T (2009) MicroRNAs with a nucleolar location. *Rna* 15:1705–1715.

[135] Sharma NR et al. (2016) Cell type- and tissue contextdependent nuclear distribution of human Ago2. *Journal of Biological Chemistry* 291:2302–2309.

[136] Kalantari R et al. (2016) Stable association of RNAi machinery is conserved between the cytoplasm and nucleus of human cells. *RNA (New York, N.Y.)* 22:1085–98.

[137] Morris KV (2004) Small Interfering RNA-Induced Transcriptional Gene Silencing in Human Cells. *Science (New York, N.Y.)* 305:1289–1292.

[138] Castanotto D et al. (2005) Short hairpin RNA-directed cytosine (CpG) methylation of the RASSF1A gene promoter in HeLa cells. *Molecular Therapy* 12:179–183.

[139] Ting AH, Schuebel KE, Herman JG, Baylin SB (2005) Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nature genetics* 37:906–910.

[140] Janowski Ba et al. (2006) Involvement of AGO1 and AGO2 in mammalian transcriptional silencing. *Nature structural & molecular biology* 13:787–792.

[141] Kim DH, SÃtrom P, SnÃve O, Rossi JJ (2008) MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proceedings of the National Academy of Sciences* 105:16230–16235.

[142] Younger ST, Corey DR (2011) Transcriptional gene silencing in mammalian cells by miRNA mimics that target gene promoters. *Nucleic Acids Research* 39:5682–5691.

[143] Fabiano AJ, Qiu J (2015) Post-stereotactic radiosurgery brain metastases: A review. *Journal of Neurosurgical Sciences* 59:157–167.

[144] Li LC et al. (2006) Small dsRNAs induce transcriptional activation in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 103:17337–42.

[145] Huang V et al. (2013) Ago1 Interacts with RNA Polymerase II and Binds to the Promoters of Actively Transcribed Genes in Human Cancer Cells. *PLoS Genetics* 9:e1003821.

[146] Gagnon K, Li L, Chu Y, Janowski B, Corey D (2014) RNAi factors are present and active in human cell nuclei. *Cell Reports* 6:211–221.

[147] Carissimi C et al. (2015) ARGONAUTE2 cooperates with SWI/SNF complex to determine nucleosome occupancy at human Transcription Start Sites. *Nucleic acids research* 43:1498–512.

[148] Cho S, Park JS, Kang YK (2014) AGO2 and SETDB1 cooperate in promoter-targeted transcriptional silencing of the androgen receptor gene. *Nucleic Acids Research* 42:13545–13556.

[149] Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ (2014) R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* 516:436–9.

[150] Buggiano V et al. (2009) Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nature Structural & Molecular Biology* 16:717–725.

[151] Ameyar-Zazoua M et al. (2012) Argonaute proteins couple chromatin silencing to alternative splicing. *Nature Structural & Molecular Biology* 19:998–1004.

[152] Taliaferro JM et al. (2013) Two new and distinct roles for Drosophila Argonaute-2 in the nucleus: Alternative pre-mRNA splicing and transcriptional repression. *Genes and Development* 27:378–389.

[153] Alló M et al. (2014) Argonaute-1 binds transcriptional enhancers and controls constitutive and alternative splicing in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 111:15622–9.

[154] Agirre E et al. (2015) A chromatin code for alternative splicing involving a putative association between CTCF and HP1α proteins. *BMC biology* 13:31.

[155] Wei W et al. (2012) A role for small RNAs in DNA double-strand break repair. *Cell* 149:101–112.

[156] Francia S et al. (2012) Site-specific DICER and DROSHA RNA products control the DNA-damage response. *Nature* 488:231–5.

[157] Gao M et al. (2014) Ago2 facilitates Rad51 recruitment and DNA double-strand break repair by homologous recombination. *Cell Research* 24:532–541.

[158] Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* 7:e30733.

[159] Hansen TB et al. (2013) Natural RNA circles function as efficient microRNA sponges. TL - 495. *Nature* 495 VN -:384–388.

[160] Memczak S et al. (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. TL - 495. *Nature* 495 VN -:333–338.

[161] Guarnerio J et al. (2016) Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations. *Cell* 165:289–302.

[162] Guo JU, Agarwal V, Guo H, Bartel DP (2014) Expanded identification and characterization of mammalian circular RNAs. *Genome biology* 15:409.

[163] Denzler R, Agarwal V, Stefano J, Bartel DP, Stoffel M (2014) Assessing the ceRNA Hypothesis with Quantitative Measurements of miRNA and Target Abundance. *Molecular Cell* 54:766–776.

[164] Bosson AD, Zamudio JR, Sharp PA (2014) Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Molecular Cell* 56:347–359.

[165] Mayya VK, Duchaine TF (2015) On the availability of microRNA-induced silencing complexes, saturation of microRNA-binding sites and stoichiometry. *Nucleic Acids Research* 43:7556–7565.

[166] Yuan Y et al. (2015) Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. *Proceedings of the National Academy of Sciences of the United States of America* 112:3158–63.

[167] Martirosyan A, Figliuzzi M, Marinari E, De Martino A (2016) Probing the Limits to MicroRNA-Mediated Control of Gene Expression. *PLoS Computational Biology* 12:e1004715.

[168] Kumar MS et al. (2014) HMGA2 functions as a competing endogenous RNA to promote lung cancer progression. *Nature* 505:212–7.

[169] Karreth FA et al. (2015) The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell* 161:319–332.

[170] Thomson DW, Dinger ME (2016) Endogenous microRNA sponges: evidence and controversy. *Nature Reviews Genetics* 17:272–283.

[171] Tan JY et al. (2015) Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. *Genome Research* 125:655–666.

[172] Militello G et al. (2016) Screening and validation of lncRNAs and circRNAs as miRNA sponges. *Briefings in Bioinformatics* p. bbw053.

[173] Iwai N, Naraba H (2005) Polymorphisms in human pre-miRNAs. *Biochemical and Biophysical Research Communications* 331:1439–1444.

[174] Duan R, Pak CH, Jin P (2007) Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Human Molecular Genetics* 16:1124–1131.

[175] Saunders Ma, Liang H, Li WH (2007) Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences of the United States of America* 104:3300–5.

[176] Landi D, Gemignani F, Barale R, Landi S (2008) A catalog of polymorphisms falling in microRNA-binding regions of cancer genes. *DNA and cell biology* 27:35–43.

[177] Carbonell J et al. (2012) A map of human microRNA variation uncovers unexpectedly high levels of variability. *Genome medicine* 4:62.

[178] Bhartiya D, Laddha SV, Mukhopadhyay A, Scaria V (2011) miRvar: A comprehensive database for genomic variations in microRNAs. *Human Mutation* 32:2226–2245.

[179] Liu C et al. (2012) MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics* 13:661.

[180] Zorc M et al. (2012) Catalog of microRNA seed polymorphisms in vertebrates. *PLoS ONE* 7.

[181] Bhattacharya A, Ziebarth JD, Cui Y (2014) PolymiRTS Database 3.0: Linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Research* 42:86–91.

[182] Borel C et al. (2011) Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. *Genome Research* 21:68–73.

[183] Rantalainen M et al. (2011) MicroRNA Expression in Abdominal and Gluteal Adipose Tissue Is Associated with mRNA Expression Levels and Partly Genetically Driven. *PLoS ONE* 6.

[184] Parts L et al. (2012) Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS Genetics* 8.

[185] Gamazon ER et al. (2012) Genetic architecture of microRNA expression: Implications for the transcriptome and complex traits. *American Journal of Human Genetics* 90:1046–1063.

[186] Lappalainen T et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506–11.

[187] Yang H et al. (2008) Evaluation of genetic variants in MicroRNA-related genes and risk of bladder cancer. *Cancer Research* 68:2530–2537.

[188] Hu Z et al. (2009) Common genetic variants in pre-microRNAs were associated with increased risk of breast cancer in Chinese women. *Human Mutation* 30:79–84.

[189] Merritt WM et al. (2008) Dicer, Drosha, and Outcomes in Patients with Ovarian Cancer. *New England Journal of Medicine* 359:2641–2650.

[190] Lu J et al. (2005) MicroRNA expression profiles classify human cancers. *Nature* 435:834–838.

[191] Ripke S et al. (2011) Genome-wide association study identifies five new schizophrenia loci.

[192] Saus E et al. (2010) Genetic variants and abnormal processing of pre-miR-182, a circadian clock modulator, in major depression patients with late insomnia. *Human Molecular Genetics* 19:4017–4025.

[193] de Pontual L et al. (2011) Germline deletion of the miR-1792 cluster causes skeletal and growth defects in humans. *Nature genetics* 43:1026–30.

[194] Calin GA et al. (2005) A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* 353:1793–1801.

[195] Chae YS et al. (2013) A miR-146a polymorphism (rs2910164) predicts risk of and survival from colorectal cancer. *Anticancer Research* 33:3233–3240.

[196] Upadhyaya A et al. (2015) Association of the microRNA-Single Nucleotide Polymorphism rs2910164 in miR146a with sporadic breast cancer susceptibility: A case control study. *Gene* 576:256–260.

[197] Jazdzewski K et al. (2008) Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* 105:7269–7274.

[198] Shen J, Ambrosone CB, Zhao H (2009) Novel genetic variants in microRNA genes and familial breast cancer. *International Journal of Cancer* 124:1178–82.

[199] Young SD (2014) [Under Review]. *American journal of public health* 75083:1–20.

[200] Mencía A et al. (2009) Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nature genetics* 41:609–613.

[201] Kuhn S et al. (2011) miR-96 regulates the progression of differentiation in mammalian cochlear inner and outer hair cells. *Proceedings of the National Academy of Sciences of the United States of America* 108:2355–2360.

[202] Travis AJ, Moody J, Helwak A, Tollervey D, Kudla G (2014) Hyb: A bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. *Methods* 65:263–273.

[203] ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

[204] McVean GA et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

[205] Derrien T et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* 22:1775–1789.

[206] Smit A, Hubley R, Green P (2015) RepeatMasker Open-4.0.

[207] Forrest ARR et al. (2014) A promoter-level mammalian expression atlas. *Nature* 507:462–470.

[208] Sherry ST et al. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29:308–11.

[209] Lek M et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–91.

[210] Gudbjartsson DF et al. (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* 47:435–444.

[211] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359.

[212] Kim D, Salzberg SL (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology* 12:R72.

[213] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215:403–10.

[214] Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods in molecular biology (Clifton, N.J.)* 453:3–31.

[215] Dobin A et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.

[216] Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34:525–527.

[217] Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15:550.

[218] Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

[219] Li H et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

[220] Pippucci T (2014) H3M2: detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics (Oxford, England)* 30:2852–2859.

[221] R Core Team (2016) R: A Language and Environment for Statistical Computing.

[222] Hadley Wickham (2009) *ggplot2: elegant graphics for data analysis*. (Springer New York).

[223] M D, Srinivasan A (2016) data.table: Extension of Data.frame.

[224] Shabalin AA (2012) Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28:1353–1358.

[225] Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8:25.

[226] Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111.

[227] Carninci P et al. (2005) The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)* 309:1559–63.

[228] Grosswendt S et al. (2014) Unambiguous Identification of miRNA: Target site interactions by different types of ligation reactions. *Molecular Cell* 54:1042–1054.

[229] Lorenz R et al. (2011) {ViennaRNA} Package 2.0. *Algorithms for Molecular Biology* 6:26.

[230] Pethick J (2015) Ph.D. thesis (The University Of Edinburgh).

[231] Bayne EH, Allshire RC (2005) RNA-directed transcriptional gene silencing in mammals. *Trends in Genetics* 21:370–373.

[232] Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften* 131:281–5.

[233] Breiman L (2001) Random forests. *Machine Learning* 45:5–32.

[234] Janitza S, Strobl C, Boulesteix AL (2013) An AUC-based permutation variable importance measure for random forests. *BMC bioinformatics* 14:119.

[235] Rutenberg-Schoenberg M, Sexton AN, Simon MD (2016) The Properties of Long Noncoding RNAs That Regulate Chromatin. *Annual review of genomics and human genetics* 17:69–94.

[236] Kanamori-Katayama M et al. (2011) Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Research* 21:1150–1159.

[237] Young RS et al. (2015) The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Research* 25:1546–1557.

[238] Jiang X et al. (2014) Expression of obesity-related miR-1908 in human adipocytes is regulated by adipokines, free fatty acids and hormones. *Molecular Medicine Reports* 10:1164–1169.

[239] Ghanbari M et al. (2015) The association of common polymorphisms in miR-196a2 with waist to hip ratio and miR-1908 with serum lipid and glucose. *Obesity* 23:495–503.

[240] Civelek M et al. (2013) Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits. *Human Molecular Genetics* 22:3023–37.

[241] Tang R et al. (2015) The polymorphic terminal-loop of pre-miR-1307 binding with MBNL1 contributes to colorectal carcinogenesis via interference with Dicer1 recruitment. *Carcinogenesis* 36:867–875.

[242] Forloni M et al. (2014) miR-146a promotes the initiation and progression of melanoma by activating Notch signaling. *eLife* 3:e01460.

[243] Taganov KD, Boldin MP, Chang KJ, Baltimore D (2006) NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proceedings of the National Academy of Sciences of the United States of America* 103:12481–6.

[244] Zhao Y et al. (2014) Polymorphisms in MicroRNAs Are Associated with Survival in Non-Small Cell Lung Cancer. *Cancer Epidemiology Biomarkers & Prevention* 23:2503–2511.

[245] Fu A et al. (2014) Targetome profiling and functional genetics implicate miR-618 in lymphomagenesis. *Epigenetics* 9:730–737.

[246] Abdalla MAK, Haj-Ahmad Y (2012) Promising candidate urinary microrna biomarkers for the early detection of hepatocellular carcinoma among high-risk hepatitis C virus Egyptian patients. *Journal of Cancer* 3:19–31.

[247] Fassan M et al. (2009) MicroRNA expression profiling of male breast cancer. *Breast cancer research : BCR* 11:R58.

[248] Ghanbari M et al. (2014) A Genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. *Human Mutation* 35:1524–1531.

[249] Li Q, Chen L, Chen D, Wu X, Chen M (2015) Influence of microRNA-related polymorphisms on clinical outcomes in coronary artery disease. *American Journal of Translational Research* 7:393–400.

[250] Ventura A et al. (2008) Targeted Deletion Reveals Essential and Overlapping Functions of the miR-1792 Family of miRNA Clusters. *Cell* 132:875–886.

[251] Liu C, Zhang Y, Chen H, Jiang L, Xiao D (2016) Function analysis of rs9589207 polymorphism in miR-92a in gastric cancer. *Tumor Biology* 37:4439–4444.

[252] Hansen TH et al. (2015) Impact of PTBP1 rs11085226 on glucose-stimulated insulin release in adult Danes. *BMC Medical Genetics* 16:17.

[253] Li YI et al. (2016) RNA splicing is a primary link between genetic variation and disease. *Science (New York, N.Y.)* 352:600–4.

[254] Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 4:1073–1081.

[255] Adzhubei IA et al. (2010) A method and server for predicting damaging missense mutations. *Nature Methods* 7:248–249.

[256] McKenna A et al. (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297–1303.

[257] Levy S et al. (2007) The diploid genome sequence of an individual human. *PLoS Biology* 5:2113–2144.

[258] Han YC et al. (2015) An allelic series of miR-1792mutant mice uncovers functional specialization and cooperation among members of a microRNA polycistron. *Nature Genetics* 47:766–775.

[259] Du P, Wang L, Sliz P, Gregory RI (2015) A Biogenesis Step Upstream of Microprocessor Controls miR-1792 Expression. *Cell* 162:885–899.

[260] Purcell S et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81:559–575.

[261] Purcell SM et al. (2014) A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506:185–190.

[262] O'Driscoll M, Ruiz-Perez VL, Woods CG, Jeggo PA, Goodship JA (2003) A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. *Nature Genetics* 33:497–501.

[263] Ellis NA et al. (1995) The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* 83:655–666.

[264] Shaheen R et al. (2014) Genomic analysis of primordial dwarfism reveals novel disease genes. *Genome Research* 24:291–299.

[265] Murray JE et al. (2014) Extreme Growth Failure is a Common Presentation of Ligase IV Deficiency. *Human Mutation* 35:76–85.

[266] Matsumoto Y et al. (2011) Two unrelated patients with MRE11A mutations and Nijmegen breakage syndrome-like severe microcephaly. *DNA Repair* 10:314–321.

[267] Shen J et al. (2010) Mutations in PNKP cause microcephaly, seizures and defects in DNA repair. *Nature genetics* 42:245–9.

[268] Nicholas AK et al. (2009) The molecular landscape of ASPM mutations in primary microcephaly. *Journal of medical genetics* 46:249–53.

[269] Rauch A et al. (2008) Mutations in the pericentrin (PCNT) gene cause primordial dwarfism. *Science (New York, N.Y.)* 319:816–819.

[270] Guernsey DL et al. (2010) Mutations in centrosomal protein CEP152 in primary microcephaly families linked to MCPH4. *American Journal of Human Genetics* 87:40–51.

[271] Waters AM et al. (2015) The kinetochore protein, CENPF, is mutated in human ciliopathy and microcephaly phenotypes. *Journal of medical genetics* 52:147–56.

[272] Al-Dosari MS, Shaheen R, Colak D, Alkuraya FS (2010) Novel CENPJ mutation causes Seckel syndrome. *Journal of medical genetics* 47:411–414.

[273] Bond J et al. (2005) A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nature genetics* 37:353–5.

[274] Wood JL, Liang Y, Li K, Chen J (2008) Microcephalin/MCPH1 associates with the condensin II complex to function in homologous recombination repair. *Journal of Biological Chemistry* 283:29586–29592.

[275] Floriot S et al. (2015) C-Nap1 mutation affects centriole cohesion and is associated with a Seckel-like syndrome in cattle. *Nature communications* 6:6894.

[276] Bicknell LS et al. (2011) Mutations in ORC1, encoding the largest subunit of the origin recognition complex, cause microcephalic primordial dwarfism resembling Meier-Gorlin syndrome. *Nature genetics* 43:350–355.

[277] Bicknell LS et al. (2011) Mutations in the pre-replication complex cause Meier-Gorlin syndrome. *Nature genetics* 43:356–9.

[278] Guernsey DL et al. (2011) Mutations in origin recognition complex gene ORC4 cause Meier-Gorlin syndrome. *Nature genetics* 43:360–364.

[279] Fenwick A et al. (2016) Mutations in CDC45, Encoding an Essential Component of the Pre-initiation Complex, Cause Meier-Gorlin Syndrome and Craniosynostosis. *The American Journal of Human Genetics* 99:125–138.

[280] Edery P et al. (2011) Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. *Science (New York, N.Y.)* 332:240–3.

[281] Yamamoto S et al. (2014) A drosophila genetic resource of mutants to study mechanisms underlying human genetic diseases. *Cell* 159:200–14.

[282] Ng D et al. (2004) Oculofaciocardiodental and Lenz microphthalmia syndromes result from distinct classes of mutations in BCOR. *Nature genetics* 36:411–416.

[283] Marzuillo P et al. (2013) Novel cAMP binding protein-BP (CREBBP) mutation in a girl with Rubinstein-Taybi syndrome, GH deficiency, Arnold Chiari malformation and pituitary hypoplasia. *BMC medical genetics* 14:28.

[284] Ji J et al. (2015) DYRK1A haploinsufficiency causes a new recognizable syndrome with microcephaly, intellectual disability, speech impairment, and distinct facies. *European journal of human genetics : EJHG* 23:1473–1481.

[285] Troelstra C, Gool AV, Wit JD, Vermeulen W, Bootsma D (1992) ERCCG , a Member of a Subfamily of Putative Helicases , Is Involved in Cockaynes Syndrome and Preferential Repair of Active Genes. *Cell* 71:939–953.

[286] Schüle B, Oviedo A, Johnston K, Pai S, Francke U (2005) Inactivating mutations in ESCO2 cause SC phocomelia and Roberts syndrome: no phenotype-genotype correlation. *American journal of human genetics* 77:1117–1128.

[287] Kortüm F et al. (2011) The core FOXG1 syndrome phenotype consists of postnatal microcephaly, severe mental retardation, absent language, dyskinesia, and corpus callosum hypogenesis. *Journal of medical genetics* 48:396–406.

[288] Ostergaard P et al. (2012) Mutations in KIF11 cause autosomal-dominant microcephaly variably associated with congenital lymphedema and chorioretinopathy. *American Journal of Human Genetics* 90:356–362.

[289] Risheg H et al. (2007) A recurrent mutation in MED12 leading to R961W causes Opitz-Kaveggia syndrome. *Nat Genet* 39:451–453.

[290] Alston CL et al. (2016) A recurrent mitochondrial p.Trp22Arg NDUFB3 variant causes a distinctive facial appearance, short stature and a mild biochemical and clinical phenotype. *Journal of Medical Genetics* pp. jmedgenet–2015–103576.

[291] Abbasi-Moheb L et al. (2012) Mutations in NSUN2 cause autosomal- Recessive intellectual disability. *American Journal of Human Genetics* 90:847–855.

[292] Dyment DA et al. (2013) Mutations in PIK3R1 cause SHORT syndrome. *American Journal of Human Genetics* 93:158–166.

[293] Martin CA et al. (2014) Mutations in PLK4, encoding a master regulator of centriole biogenesis, cause microcephaly, growth failure and retinopathy. *Nature genetics* 46:1283–92.

[294] Senderek J et al. (2005) Mutations in SIL1 cause Marinesco-Sjögren syndrome, a cerebellar ataxia with cataract and myopathy. *Nature genetics* 37:1312–4.

[295] Boerkoel CF et al. (2002) Mutant chromatin remodeling protein SMARCAL1 causes Schimke immuno-osseous dysplasia. *Nature genetics* 30:215–220.

[296] Hood RL et al. (2012) Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. *American Journal of Human Genetics* 90:308–313.

[297] Scheidecker S et al. (2015) Mutations in TUBGCP4 alter microtubule organization via the γ-tubulin ring complex in autosomal-recessive microcephaly with chorioretinopathy. *American journal of human genetics* 96:666–74.

[298] Balikova I et al. (2009) Deletions in the VPS13B (COH1) gene as a cause of Cohen syndrome. *Human Mutation* 30:E845–E854.

[299] Verma AS, Fitzpatrick DR (2007) Anophthalmia and microphthalmia. *Orphanet journal of rare diseases* 2:47.

[300] Gerth-Kahlert C et al. (2013) Clinical and mutation analysis of 51 probands with anophthalmia and/or severe microphthalmia from a single center. *Molecular genetics & genomic medicine* 1:15–31.

[301] Ferda Percin E et al. (2000) Human microphthalmia associated with mutations in the retinal homeobox gene CHX10. *Nature genetics* 25:397–401.

[302] Voronina VA et al. (2004) Mutations in the human RAX homeobox gene in a patient with anophthalmia and sclerocornea. *Human Molecular Genetics* 13:315–322.

[303] Glaser T et al. (1994) PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. *Nature genetics* 7:463–471.

[304] Fantes J et al. (2003) Mutations in SOX2 cause anophthalmia. *Nature Genetics* 33:461–463.

[305] Ragge NK et al. (2005) Heterozygous mutations of OTX2 cause severe ocular malformations. *American journal of human genetics* 76:1008–1022.

[306] Aligianis Ia et al. (2005) Mutations of the catalytic subunit of RAB3GAP cause Warburg Micro syndrome. *Nature genetics* 37:221–223.

[307] Pasutto F et al. (2007) Mutations in STRA6 cause a broad spectrum of malformations including anophthalmia, congenital heart defects, diaphragmatic hernia, alveolar capillary dysplasia, lung hypoplasia, and mental retardation. *American journal of human genetics* 80:550–560.

[308] Reis LM et al. (2011) BMP4 loss-of-function mutations in developmental eye disorders including SHORT syndrome. *Human Genetics* 130:495–504.

[309] Rainger J et al. (2011) Loss of the BMP antagonist, SMOC-1, causes Ophthalmo-acromelic (Waardenburg anophthalmia) syndrome in humans and mice. *PLoS Genetics* 7:e1002114.

[310] Semina EV, Brownell I, Mintz-Hittner Ha, Murray JC, Jamrich M (2001) Mutations in the human forkhead transcription factor FOXE3 associated with anterior segment ocular dysgenesis and cataracts. *Human molecular genetics* 10:231–6.

[311] Bem D et al. (2011) Loss-of-function mutations in RAB18 cause Warburg micro syndrome. *American Journal of Human Genetics* 88:499–507.

[312] Liegel RP et al. (2013) Loss-of-function mutations in TBC1d20 cause cataracts and male infertility in blind sterile mice and Warburg micro syndrome in humans. *American Journal of Human Genetics* 93:1001–1014.

[313] Kevelam SH et al. (2015) Recessive ITPA mutations cause an early infantile encephalopathy. *Annals of Neurology* 78:649–658.

[314] Morin PJ (1997) Activation of beta -Catenin-Tcf Signaling in Colon Cancer by Mutations in beta -Catenin or APC. *Science* 275:1787–1790.

[315] Gemignani F et al. (2004) A TP53 polymorphism is associated with increased risk of colorectal cancer and with reduced levels of TP53 mRNA. *Oncogene* 23:1954–1956.

[316] Vogelstein B et al. (1988) Genetic Alterations During Colorectal-Tumour Development. *The New England Journal of Medicine* 319:525–532.

[317] Philp AJ et al. (2001) The phosphatidylinositol 3-kinase p85α gene is an oncogene in human ovarian and colon tumors. *Cancer Research* 61:7426–7429.

[318] Mao JH et al. (2004) Fbxw7/Cdc4 is a p53-dependent, haploinsufficient tumour suppressor gene. *Nature* 432:775–779.

[319] Woodford-Richens KL et al. (2001) SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. *Proceedings of the National Academy of Sciences of the United States of America* 98:9719–23.

[320] Bass AJ et al. (2011) Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature genetics* 43:964–8.

[321] Davies H et al. (2002) Mutations of the BRAF gene in human cancer. *Nature* 417:949–954.

[322] Takagi Y et al. (1998) Somatic alterations of the SMAD-2 gene in human colorectal cancers. *British journal of cancer* 78:1152–5.

[323] Wang H et al. (2010) PCBP1 Suppresses the Translation of Metastasis-Associated PRL-3 Phosphatase. *Cancer Cell* 18:52–62.